Chapter 7 :Q1

Kholoud Basalim
2024-11-23

Q7.1:

The number of deaths from leukemia and other cancers among survivors of the Hiroshima atom bomb are shown in Table 7.12, classified by the radiation dose received. The data refer to deaths during the period 1950–1959 among survivors who were aged 25 to 64 years in 1950 (from dataset 13 of Cox and Snell 1981, attributed to Otake 1979).

Table 7.12: *y = Number of Deaths from leukemia and other cancers classified by radiation dose received from the Hiroshima atomic bomb.*

	Radiation dose (rads)					
Deaths	0	1–9	10–49	50–99	100–199	200+
Leukemia	13	5	5	3	4	18
Other cancers	378	200	151	47	31	33
Total cancers	391	205	156	50	35	51

(a) Obtain a suitable model to describe the dose–response relationship between radiation and the proportional cancer mortality rates for leukemia.
(b) Examine how well the model describes the data.
(c) Interpret the results.

Let:

$n_i = Total\ cancers$
$y_i = Number\ of\ deaths\ from\ leukemia$
$n_i - y_i = Number\ of\ deaths\ from\ other\ cancers$
$x_i = radiation\ dose\ (lower\ limit\ of\ radiation\ dose\ interval)\ i = 1,2,\dots,N$
$N = 6\ (Number\ of\ different\ values\ of\ radiation\ dose\ x_i)$
$p_i = \frac{y_i}{n_i}\ Proportion\ of\ deaths\ from\ leukemi$

```
#Deaths by Leukemia
y<-c(13,5,5,3,4,18)
#n=Total number of deaths by cancers
n<-c(391,205,156,50,35,51)
#Deaths by other cancers
n_y<- n-y
#P=Proportion of deaths from leukemia
p=y/n
#x=radiation dose(lower limit of radiation dose interval)
x<-c(0,1,10,50,100,200)
```

Data on the table (df):

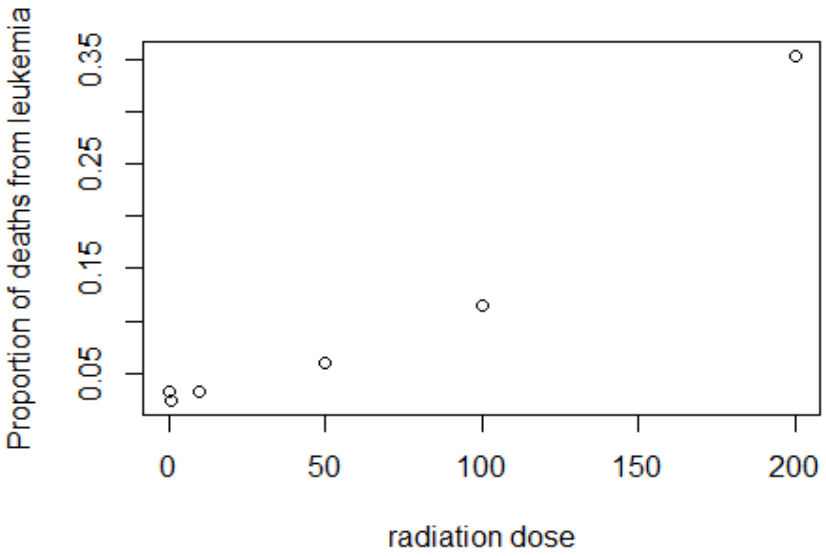
```
#but the data in table :
df<- data.frame(x,y,n_y,n,p)
df

##     x  y n_y   n          p
## 1   0 13 378 391 0.03324808
## 2   1  5 200 205 0.02439024
## 3  10  5 151 156 0.03205128
## 4  50  3  47  50 0.06000000
## 5 100  4  31  35 0.11428571
## 6 200 18  33  51 0.35294118
```

x_i	y_i	$n_i - y_i$	n_i	$p_i = \frac{y_i}{n_i}$
0	13	378	391	0.03324808
1	5	200	205	0.02439024
10	5	151	156	0.03205128
50	3	47	50	0.06
100	4	31	35	0.11428571
200	18	33	33	0.35294118

graph between x_i and p_i :

```
#plot x=radiation dose vs p=Proportion of deaths from leukemia :
plot(x,p , xlab = "radiation dose" , ylab = "Proportion of deaths from leukemia")
```



Non-linear relationship:

The relationship between the two variables is not a simple linear one (i.e., it cannot be represented by a straight line). Increasing mortality rate with increasing value X : Generally, as the value of the independent variable (X) increases, the percentage of deaths due to leukemia (P) also increases. This indicates a positive correlation between the two variables.

Because the study represents the probability of a binary outcome (in this case, whether a person dies of leukemia or other cancer) based on one explanatory variable (X), The suggested model is the logistic regression model given by :

$$\begin{cases} \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i \quad , i = 1,2, \dots . N \\ y_i \sim Bin(n_i, \pi_i) \end{cases}$$

```
model<-glm(p~x ,family = binomial("logit"),weights = n)
summary(model)

##
## Call:
## glm(formula = p ~ x, family = binomial("logit"), weights = n)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.488973   0.204062 -17.098  < 2e-16 ***
## x            0.014410   0.001817   7.932 2.15e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 54.35089  on 5  degrees of freedom
## Residual deviance:  0.43206  on 4  degrees of freedom
## AIC: 26.097
##
## Number of Fisher Scoring iterations: 4
```

The final estimate of β is : $b_1 = -3.488973 \ \ and \ \ b_2 = 0.014410$

$$\text{Model} = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = b_1 + b_2 x_i \ \ >>>>>>> \ \ln\left(\frac{\pi_i}{1-\pi_i}\right) = -3.489 + 0.01441 x_i$$

The Standard error of b is : $s.e(b_{1)} = 0.204062 \ \ and \ \ s.e(\ b_2) = 0.001817$

$$\boldsymbol{b_1 \pm Z_{1-\frac{\alpha}{2}} * S.E(b_1)} \gg \ -3.488973 \pm 1.96 * 0.204062 \gg (-3.889 \ , -3.089)$$
$$\boldsymbol{b_2 \pm Z_{1-\frac{\alpha}{2}} * S.E(b_2)} \gg 0.014410 \pm 1.96 * 0.001817 \gg (0.01085 \ , 0.01797)$$

***Deviance*:**

$D_0 = 54.35089 \quad df = N - q = 6 - 1 = 5$

$D_1 = 0.43206 \quad df = N - p = 6 - 2 = 4$

A $(1-\alpha)100\%$ confidence interval for β_1 and β_2 :

```
confint.default(model ,level = 0.95)
```

```
##                    2.5 %      97.5 %
## (Intercept) -3.88892808 -3.08901798
## x            0.01084967  0.01797081
```

The variance-covariance matrix of b is:

$$\tau^{-1} = cov(\hat{\beta}) = cov(b) = \begin{bmatrix} \mathrm{var}(b_1) & \mathrm{cov}(b_1, b_2) \\ \mathrm{cov}(b_1, b_2) & \mathrm{var}(b_2) \end{bmatrix}$$

```
vcov(model)
```

```
##              (Intercept)            x
## (Intercept)  0.041641483 -2.35714e-04
## x           -0.000235714  3.30022e-06
```

The information matrix is:

$$\tau = cov(U) = \begin{bmatrix} \mathrm{var}(U_1) & \mathrm{cov}(U_1, U_2) \\ \mathrm{cov}(U_1, U_2) & \mathrm{var}(U_2) \end{bmatrix}$$

```
Tau<-solve(vcov(model))
Tau
```

```
##             (Intercept)          x
## (Intercept)    40.31297   2879.303
## x            2879.30263 508660.621
```

Odds Ratio (OR):

$$OR = e^{\beta_2} = e^{0.014410} = 1.0145$$

```
exp(model$coefficients[2])
```

```
##        x
## 1.014515
```

95% C.I of OR:

$$e^{b_2 - Z_{1-\frac{\alpha}{2}} * S.E(b_2)} < OR < e^{b_2 + Z_{1-\frac{\alpha}{2}} * S.E(b_2)}$$

$$e^{0.01085} < OR < e^{0.01797}$$

$$1.0109 < OR < 1.01813$$

```
exp(confint.default(model))
```

```
##                  2.5 %     97.5 %
## (Intercept) 0.02046727 0.04554666
## x           1.01090874 1.01813326
```

The estimate values of the probabilities ($\hat{\pi}_i$):

$$\hat{\pi}_i = \frac{e^{b_1 + b_2 x_i}}{1 + e^{b_1 + b_2 x_i}} = \frac{e^{-3.489 + 0.01441 x_i}}{1 + e^{-3.489 + 0.01441 x_i}}$$

```
#The estimates of probabilities (pi_hat) :
pi_hat<- fitted.values(model)
pi_hat
```

```
##          1          2          3          4          5          6
## 0.02962762 0.03004473 0.03406353 0.05905247 0.11425978 0.35276092
```

The fitted values of (y_i) are:

$$\boldsymbol{y_i = n_i * \hat{\pi}_i}$$

```
#Since E(Yi)=ni*pi, the fitted value of Yi (yhat):
yhat<- n*pi_hat
yhat
```

```
##         1         2         3         4         5         6
## 11.584398  6.159169  5.313911  2.952623  3.999092 17.990807
```

Goodness of fit Tests:

Hypothesis:

$$H_0: Model\ fit\ data\ well \quad vs \quad H_1: Model\ dose\ not\ fit\ data\ well$$

Test statistics:

By Deviance statistic : D=0.4321 and By Pearson Chi-squared statistics : $X^2 = 0.43$

Critical Value:

The critical value is $\chi^2_{\alpha,(N-p)} = \chi^2_{0.05,(6-2)} = \chi^2_{0.05,(4)} = 9.48773$

Decision:

Since $D = 0.4321 < \chi^2_{\alpha,(N-p)}$, we conclude that the model fits the data well.
Since $X^2 = 0.42 < \chi^2_{\alpha,(N-p)}$, we conclude that the model fits the data well.

1- Deviance (D):

```
#Test statistics (Deviance)
D<- deviance(model)
D

## [1] 0.4320565

#df=N-p , p=# of parameters =2
df_D=6-2
df_D

## [1] 4

#Critacl Value:
chi_table<- qchisq(1-0.05,df_D)
chi_table

## [1] 9.487729

#Decision:
if(D>chi_table)
{print("Reject H0")
}else{
  print("Do not Reject H0 ")
}

## [1] "Do not Reject H0 "
```

or we can find Test statistics (Deviance) by using Deviance residual

```
#The Deviance Residuals:
Deviance_Residuals<-residuals(model, type = "deviance")
Deviance_Residuals

##             1             2             3             4             5
##  0.4142808205 -0.4899417477 -0.1399058934  0.0283527664  0.0004823665
##             6
##  0.0026939960
```

```r
#test statistic (Deviance):
D_by_Residual<- sum(Deviance_Residuals^2)
D_by_Residual
```

```
## [1] 0.4320565
```

We conclude that the model is adequate for fitting the data based on the deviance

2- Pearson Chi-squared Statistic:

```r
#The Pearson (Chi-Squared) Residuals:
Pearson_Residuals<- residuals(model, type = "pearson")
Pearson_Residuals
```

```
##             1             2             3             4             5
##  0.4222166621 -0.4742527478 -0.1385560523  0.0284235278  0.0004823824
##             6
##  0.0026941004
```

```r
#test statistic (Pearson Chi-Squar):
chi_square<- sum(Pearson_Residuals^2)
#Critacl Value: df=N-p
chi_table<- qchisq(1-0.05,4)
```

We conclude that the model is adequate for fitting the data based on the Pearson Chi-squared statistics.

pseudo R^2:

```r
R_sq<- (model$null.deviance-model$deviance)/(model$null.deviance) ) *100
R_sq
```

```
## [1] 99.20506
```

$$R^2 = 1 - \frac{D_1}{D_0} = 0.99205 \quad \ggg R^2 = 99.20\%$$

The value ($pseudo\ R^2$ = 0.99) indicates that the model of interest provides good fit for the data

*R-Squared, ranges from 0 to 1, with higher values indicating a better model fit.

The following table contains some calculations:

```
df1<-data.frame(x,n,y,p,pi_hat,yhat ,Deviance_Residuals ,Pearson_Residuals)
df1
```

```
##     x   n  y          p     pi_hat      yhat Deviance_Residuals
## 1   0 391 13 0.03324808 0.02962762 11.584398       0.4142808205
## 2   1 205  5 0.02439024 0.03004473  6.159169      -0.4899417477
## 3  10 156  5 0.03205128 0.03406353  5.313911      -0.1399058934
## 4  50  50  3 0.06000000 0.05905247  2.952623       0.0283527664
## 5 100  35  4 0.11428571 0.11425978  3.999092       0.0004823665
## 6 200  51 18 0.35294118 0.35276092 17.990807       0.0026939960
##   Pearson_Residuals
## 1      0.4222166621
## 2     -0.4742527478
## 3     -0.1385560523
## 4      0.0284235278
## 5      0.0004823824
## 6      0.0026941004
```

x_i	n_i	y_i	$P_i = \frac{y_i}{n_i}$	$\hat{\pi}_i$	e_{Di} Deviance Residual	e_{Pi} Pearson Residual	$\hat{y}_i$
0	391	13	0.0332	0.029628	0.4143	0.4222	11.584
1	205	5	0.0244	0.030045	-0.4899	-0.4743	6.159
10	156	5	0.0321	0.034064	-0.1399	-0.1386	5.314
50	50	3	0.0600	0.059052	0.0284	0.0284	2.953
100	35	4	0.1143	0.114260	0.0005	0.0005	3.999
200	51	18	0.3529	0.352761	0.0027	0.0027	17.991
Sum	$\sum_{=888} n_i$	$\sum_{=48} y_i$			$D = \sum_{k=1}^{m} e_{Dk}^2 = 0.432057$	$X^2 = \sum_{k=1}^{m} e_{Pk}^2 = 0.423196$	$\sum_{=48} \hat{y}_i$

Graphs: Visualization of the fitted curve:

Plot x with p_i and $\widehat{p_i}$:

The following figures show:

(1) The observed proportions ($p_{i=\frac{y_i}{n_i}}$) plotted against the radiation dose (x_i).

(2) The expected proportions (estimates of the probabilities) ($\hat{\pi}i$) plotted against the radiation dose (x_i).

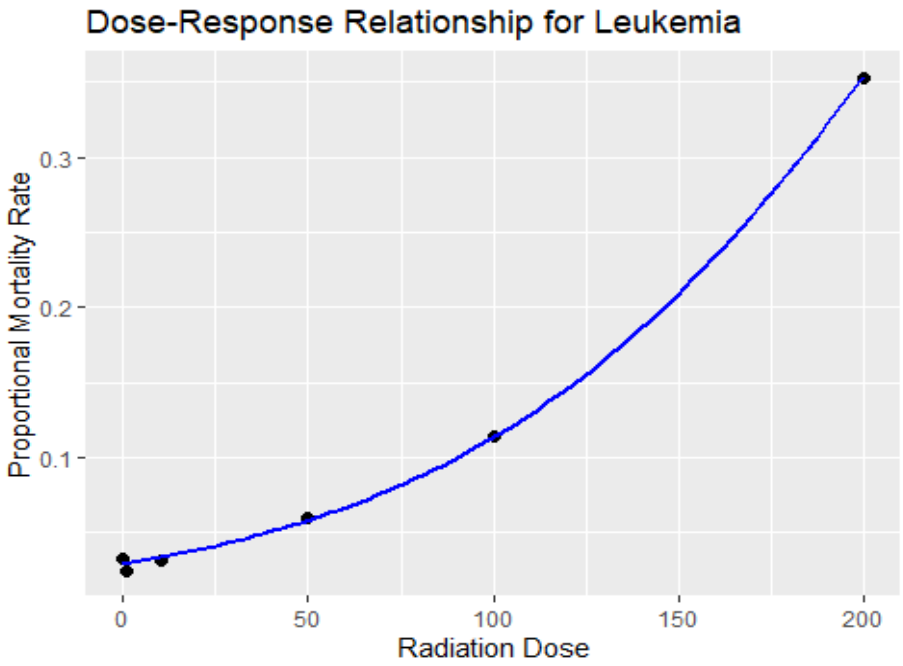
```
# Visualization of the fitted curve
#install.packages("ggplot2")
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.2
```

```
ggplot(df, aes(x = x, y = p)) +
  geom_point(size = 2) +
  stat_smooth(method = "glm", method.args = list(family = binomial(link = "logit"))
, se = FALSE, color = "blue") +
  labs(title = "Dose-Response Relationship for Leukemia",
       x = "Radiation Dose",
       y = "Proportional Mortality Rate")
```

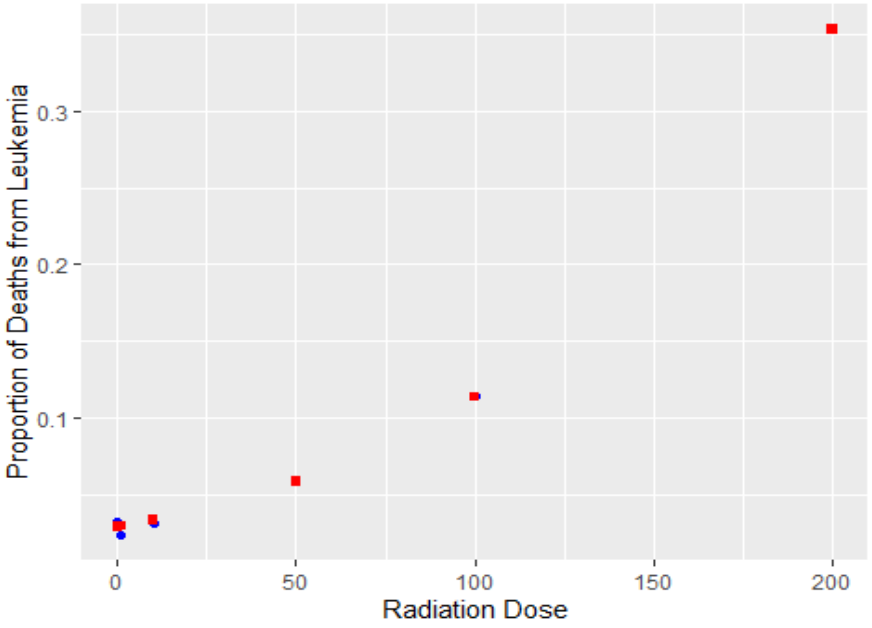
```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```



Black dots represent the observed proportions of leukaemia deaths in each radiation dose category.
The blue line represents the fitted curve estimated using a logistic regression model.

```
ggplot(data=df, aes(x = x)) +
     geom_point(aes(y = p), color="blue",shape=16) +
     geom_point(aes(y = pi_hat), color = "red",shape=15) +
     labs(x = "Radiation Dose",y = "Proportion of Deaths from Leukemia")
```



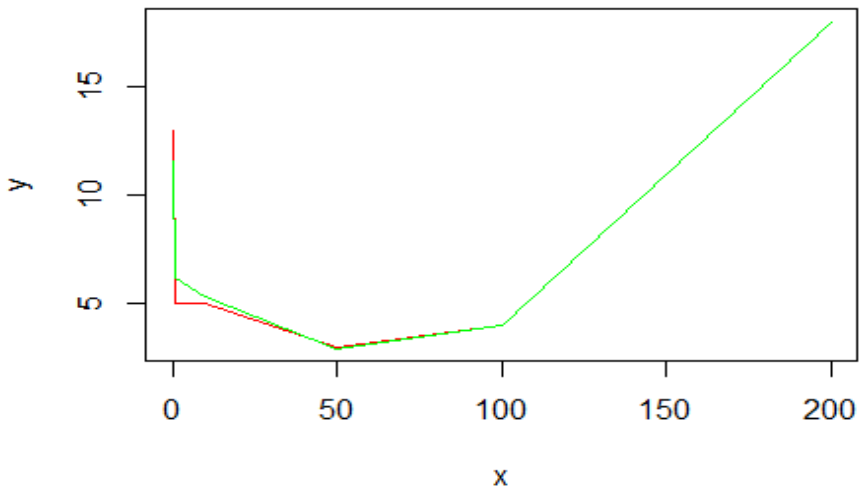
Plot x with y_i and $\hat{y}_i$:

The following figures show:

(1) The observed response (y_i) plotted against the radiation dose (x_i).

(2) The observed response ($\hat{y}_i$) plotted against the radiation dose (x_i).

```
#The observed response (yi) plotted against (xi).
#The fitted response (yhat) plotted against (xi).
plot(x,y,type="l",col="red")
lines(x,yhat,col="green")
```



OR by code:

```
#or
ggplot(data=df, aes(x = x)) +
     geom_line(aes(y = y), color="red") +
     geom_line(aes(y = pi_hat*n), color = "green") +
     labs(x = "Radiation Dose",y = "Number of Deaths ")
```