Chapter 7

Q 7.1:

Table 7.14: $y =$ 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

Let:

$n_i =$ Total cancers

$y_i =$ Number of deaths from leukemia

$n_i - y_i =$ Number of deaths from other cancers

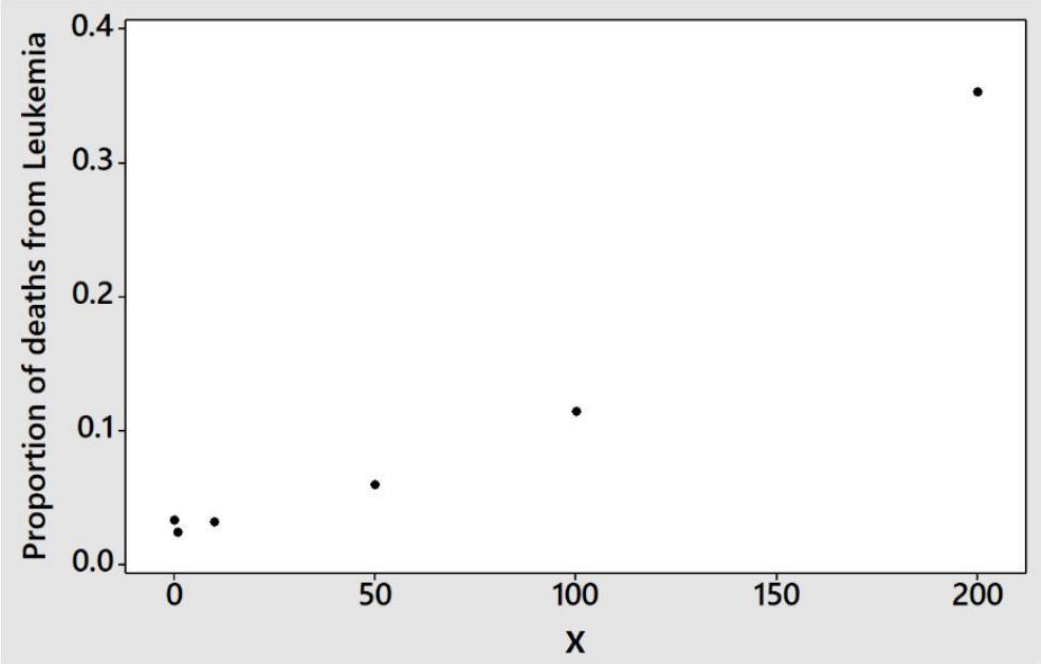
$x_i =$ radiation dose (lower limit of radiation dose interval)

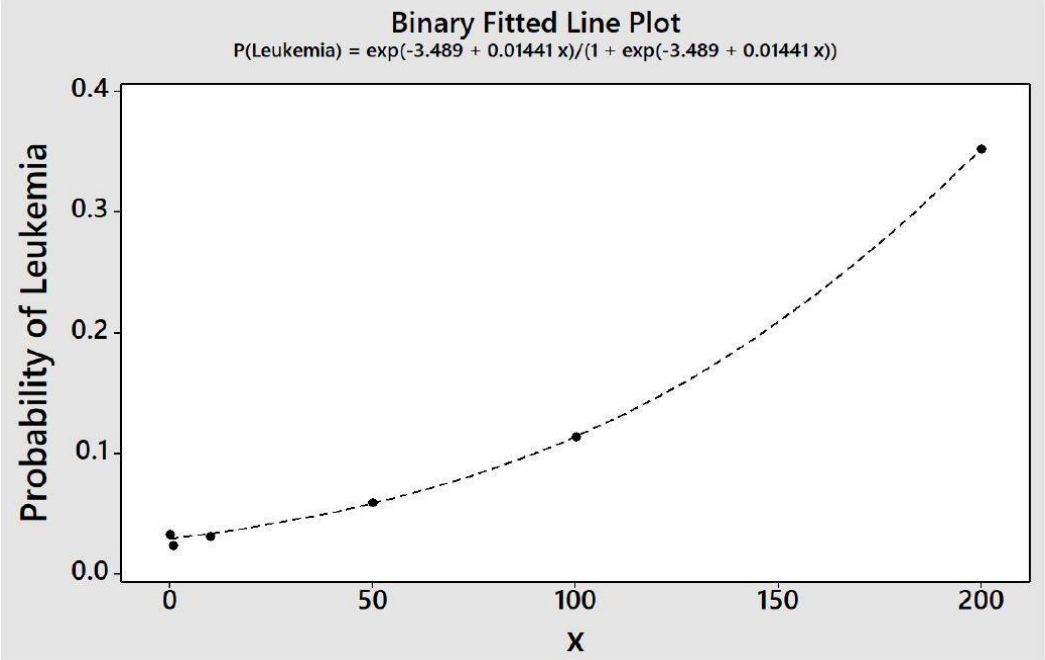
$i = 1,2, \dots, N$

$N = 6$ (Number of different values of radiation dose x_i)

$P_i = \frac{y_i}{n_i} =$ Proportion of deaths from leukemia

x_i	y_i	$n_i - y_i$	n_i	$P_i = \frac{y_i}{n_i}$
0	13	378	391	0.0332
1	5	200	205	0.0244
10	5	151	156	0.0321
50	3	47	50	0.0600
100	4	31	35	0.1143
200	18	33	51	0.3529





- The suggested model is the logistic regression model given by:

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

- Minitab Calculations:

Minitab Output:

Binary Logistic Regression: y versus x

```
Method
Link function  Logit
Rows used      6

Response Information
Variable  Value      Count  Event Name
y         Event         48  Leukemia
          Non-event    840
n         Total        888

Deviance Table
Source      DF  Seq Dev  Contribution  Adj Dev  Adj Mean  Chi-Square  P-Value
Regression   1  53.9188        99.21%  53.9188   53.9188       53.92    0.000
  x          1  53.9188        99.21%  53.9188   53.9188       53.92    0.000
Error        4   0.4321         0.79%   0.4321    0.1080
Total        5  54.3509       100.00%

Model Summary
Deviance   Deviance
    R-Sq  R-Sq(adj)     AIC
  99.21%     97.37%  323.54

Coefficients
Term         Coef  SE Coef        95% CI        Z-Value  P-Value   VIF
Constant   -3.489    0.204  ( -3.889,  -3.089)   -17.10    0.000
x         0.01441  0.00182  (0.01085, 0.01797)     7.93    0.000  1.00

Odds Ratios for Continuous Predictors
   Odds Ratio       95% CI
x      1.0145  (1.0109, 1.0181)

Regression Equation
P(Leukemia)  =  exp(Y')/(1 + exp(Y'))
Y' = -3.489 + 0.01441 x

Goodness-of-Fit Tests
Test             DF  Chi-Square  P-Value
Deviance          4        0.43    0.980
Pearson           4        0.42    0.981
Hosmer-Lemeshow   2        0.42    0.810
```

- From this output we have:

$$\hat{\beta}_1 = -3.489, \text{se}(\hat{\beta}_1) = 0.204, 95\% \text{ C.I for } \beta_1 \text{ is } (-3.889, -3.089)$$

$$\hat{\beta}_2 = 0.01441, \text{se}(\hat{\beta}_2) = 0.00182, 95\% \text{ C.I for } \beta_2 \text{ is } (0.01085, 0.01797)$$

$$\text{Deviance} : D = 0.4321 \text{ (with } df = N - p = 6 - 2 = 4)$$

Pearson Chi - Square Statstics: $X^2 = 0.4232$ (with $df = N - p = 6 - 2 = 4$)

- Variance-covariance matrix of $\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$ is:

$$\text{Cov}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} 0.0416415 & -0.0002357 \\ -0.0002357 & 0.0000033 \end{bmatrix}$$

- The fitted equation of the model is:

$$\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_1 + \hat{\beta}_2 x_i = -3.489 + 0.01441 x_i$$

- The estimates of probabilities are:

$$\hat{\pi}_i = \frac{e^{\widehat{\beta}_1 + \widehat{\beta}_2 x_i}}{1 + e^{\widehat{\beta}_1 + \widehat{\beta}_2 x_i}} = \frac{e^{-3.489 + 0.01441 x_i}}{1 + e^{-3.489 + 0.01441 x_i}}$$

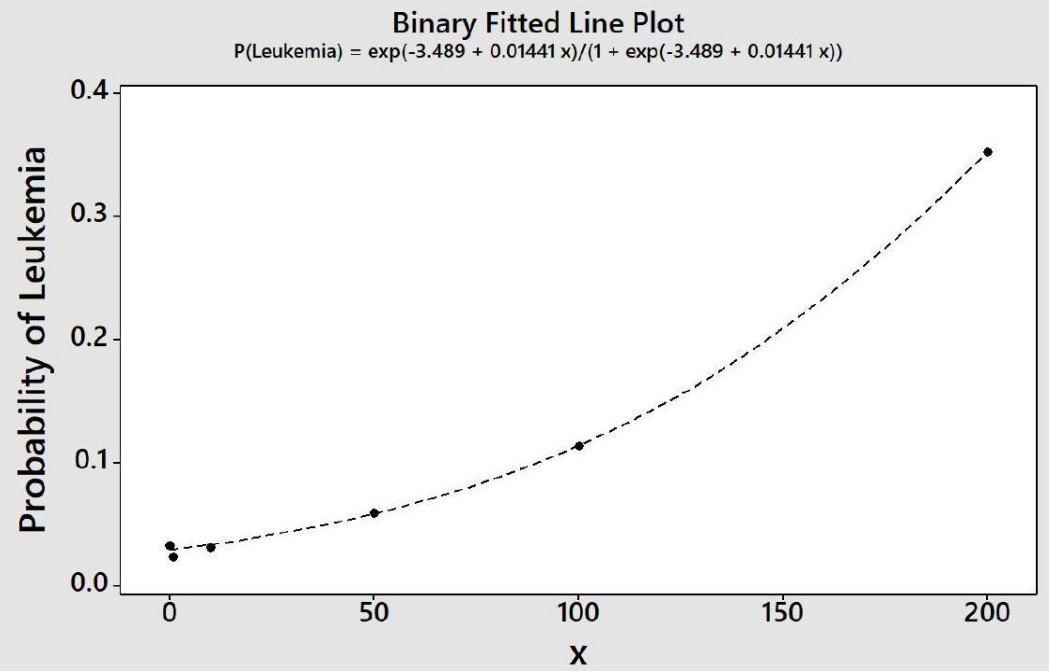
- The critical value is $\chi^2_{\alpha,(N-p)} = \chi^2_{0.05,(4)} = 9.48773$ at $\alpha = 0.05$.

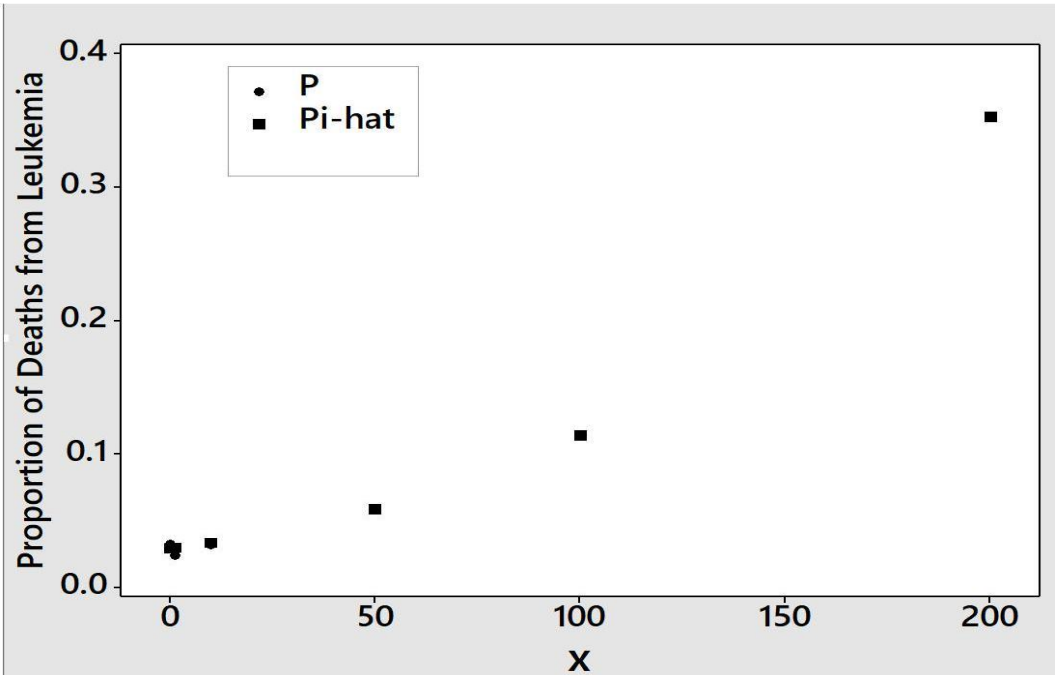
Since $D = 0.4321 < \chi^2_{0.05,(4)} = 9.48773$ (and $X^2 = 0.42 < \chi^2_{0.05,(4)} = 9.48773$), we conclude that the model fits the data well.

- The following table contains some calculations:

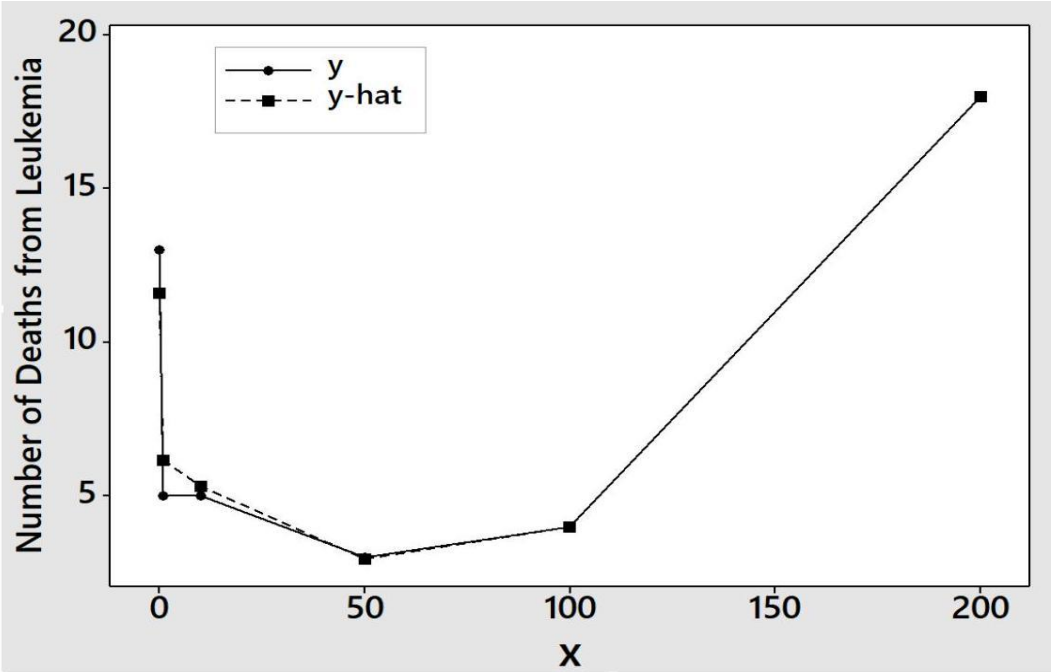
x_i	n_i	y_i	$P_i = \frac{y_i}{n_i}$	$\hat{\pi}_i$	e_{Di} Deviance Residual	e_{Pi} Pearson Residual	$\hat{y}_i$
0	391	13	0.0332	0.029628	0.4143	0.4222	11.584
1	205	5	0.0244	0.030045	-0.4899	-0.4743	6.159
10	156	5	0.0321	0.034064	-0.1399	-0.1386	5.314
50	50	3	0.0600	0.059052	0.0284	0.0284	2.953
100	35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

- 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).





- 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).



By using R:

```
#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)
```

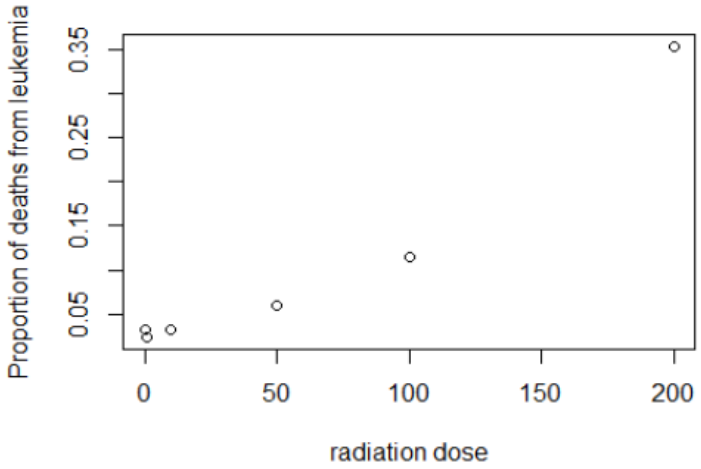
Data on the table (df):

```
#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
```

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")
```




```
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
```

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:

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

```
#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
```

```
#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:

```
#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
```

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

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

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

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
```

Graphs: Visualization of the fitted curve:

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

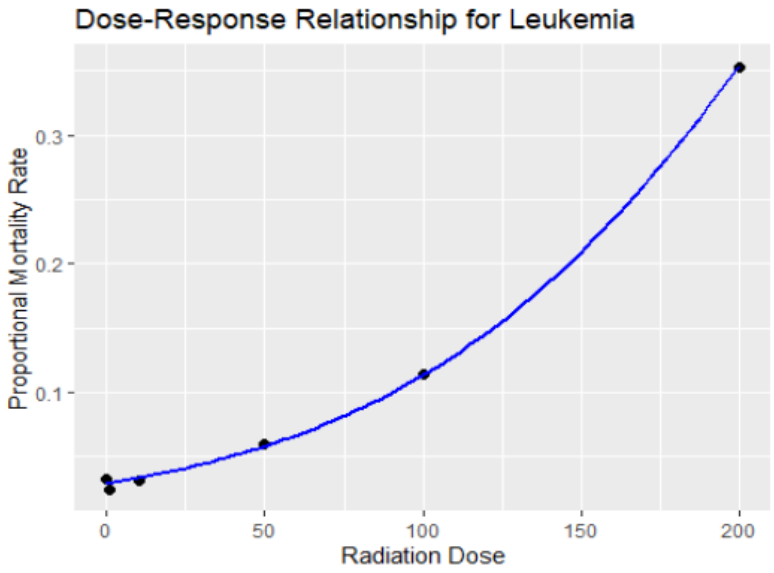
```
# Visualization of the fitted curve
#install.packages("ggplot2")
library(ggplot2)
```

```
## 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")
```

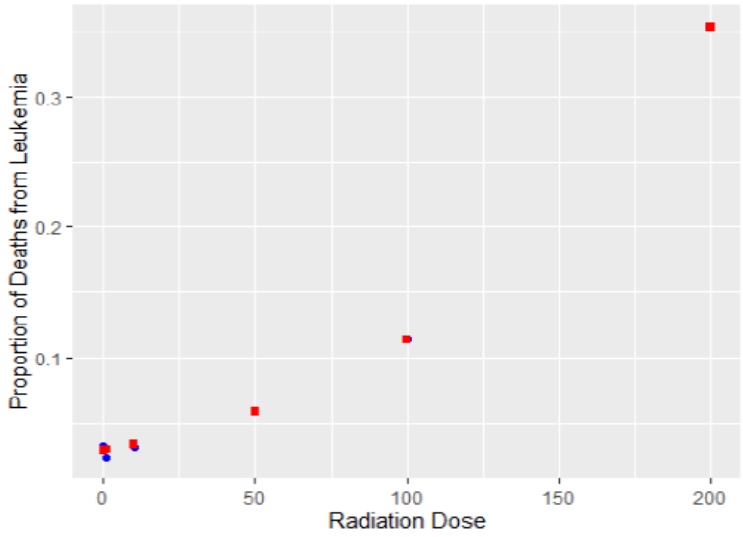
```
## `geom_smooth()` using formula = 'y ~ x'
```

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



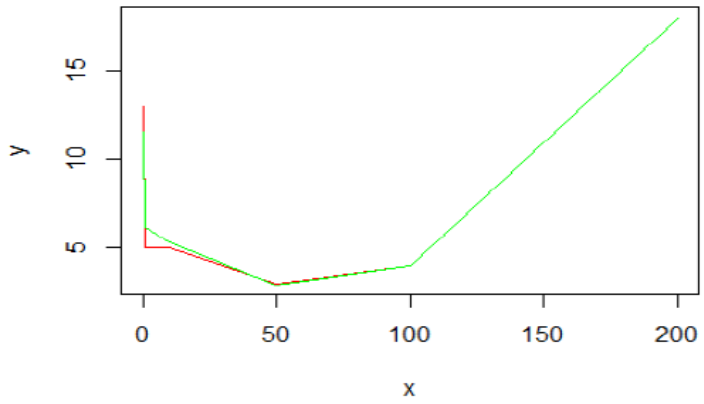
Black dots represent the observed proportions of leukaemia deaths in each radiation dose category.
The blue line represents the fitted curve estimated using a logistic regression model.

```
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 $\widehat{y_i}$:

```
#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 ")
```

Q 7.2:

Table 7.15: 2×2 table for a prospective study of exposure and disease outcome

	Diseased	Not diseased	Odds	Odds Ratio
Exposed	π_1	$1-\pi_1$	$O_1 = \frac{\pi_1}{1-\pi_1}$	
Not exposed	π_2	$1-\pi_2$	$O_2 = \frac{\pi_2}{1-\pi_2}$	$OR = \phi = \frac{O_1}{O_2}$

- The odds of disease for either exposure group is:

$$O_i = \frac{\pi_i}{1-\pi_i}; i = 1,2$$

- The odds ratio (OR) is:

$$\phi = \frac{O_1}{O_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

- The odds ratio is a measure of the relative likelihood of disease for the exposed and non-exposed groups.

(a) For the simple logistic model is:

$$\pi_i = \frac{e^{\beta_i}}{1+e^{\beta_i}}; i = 1,2 \Leftrightarrow \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_i \Leftrightarrow \frac{\pi_i}{1-\pi_i} = e^{\beta_i}$$

The odds is:

$$O_i = \frac{\pi_i}{1-\pi_i} = e^{\beta_i}$$

The odds ratio is:

$$\phi = \frac{O_1}{O_2} = \frac{e^{\beta_1}}{e^{\beta_2}} = e^{\beta_1-\beta_2}$$

If there is no difference between the exposed and not exposed groups (i.e., $\beta_1 = \beta_2 = \beta$), then the odds ratio is:

$$\phi = e^{\beta_1-\beta_2} = e^{\beta-\beta} = e^0 = 1$$

(b) Suppose that we have J age groups. Let x_j be the mean age of the j-th age group ($j = 1,2, \dots, J$), and the 2×2 contingency table for the j-th age group is:

	x_j			
	Diseased	Not diseased	Odds	Odds Ratio
Exposed	π_{1j}	$1 - \pi_{1j}$	$O_{1j} = \frac{\pi_{1j}}{1 - \pi_{1j}}$	
Not exposed	π_{2j}	$1 - \pi_{2j}$	$O_{2j} = \frac{\pi_2}{1 - \pi_{2j}}$	$\phi_j = \frac{O_{1j}}{O_{2j}}$

Consider the following logistic model:

$$\pi_{ij} = \frac{e^{\alpha_i + \beta_i x_j}}{1 + e^{\alpha_i + \beta_i x_j}} \; ; i = 1,2 \text{ and } j = 1,2, \dots, J \Leftrightarrow \ln\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \alpha_i + \beta_i x_j ; i = 1,2 \text{ and } j = 1,2, \dots, J$$

$$\Leftrightarrow \frac{\pi_{ij}}{1 - \pi_{ij}} = e^{\alpha_i + \beta_i x_j} ; i = 1,2 \text{ and } j = 1,2, \dots, J$$

The odds is for the level x_j is:

$$O_{ij} = \frac{\pi_{ij}}{1 - \pi_{ij}} = e^{\alpha_i + \beta_i x_j}$$

The odds ratio for the level x_j is:

$$\phi_j = \frac{O_{1j}}{O_{2j}} = \frac{e^{\alpha_1 + \beta_1 x_j}}{e^{\alpha_2 + \beta_2 x_j}} = e^{(\alpha_1 - \alpha_2) + (\beta_1 - \beta_2) x_j}$$

If $\beta_1 = \beta_2 = \beta$, then the odds ratio for the level x_j is:

$$\phi_j = e^{(\alpha_1 - \alpha_2)} = \phi ; j = 1,2, \dots, J$$

Thus, the odds ratios are equal for all x_j (i.e., $\phi_j = \phi =$ constant). Consequently, $\ln(\phi_j) = \ln(\phi) =$ constant for all $; j = 1,2, \dots, J$.

Q 7.3:

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

Year (X)	Sex (V)	Faculty (W)	Total (n)	Survive (Y)
1938	men	medicine	22	18
1939	men	medicine	23	16
1940	men	medicine	17	7
1941	men	medicine	25	12
1942	men	medicine	50	24
1943	men	medicine	21	16
1944	men	medicine	32	22
1945	men	medicine	14	12
1946	men	medicine	34	22
1947	men	medicine	37	28
1938	men	arts	30	16
1939	men	arts	22	13
1940	men	arts	25	11
1941	men	arts	14	12

1942	men	arts	12	8
1943	men	arts	20	11
1944	men	arts	10	4
1945	men	arts	12	4
1946	men	arts	*	*
1947	men	arts	23	13
1938	men	science	14	9
1939	men	science	12	9
1940	men	science	19	12
1941	men	science	15	12
1942	men	science	28	20
1943	men	science	21	16
1944	men	science	31	25
1945	men	science	38	32
1946	men	science	5	4
1947	men	science	31	25
1938	men	engineering	16	10
1939	men	engineering	11	7
1940	men	engineering	15	12
1941	men	engineering	9	8
1942	men	engineering	7	5
1943	men	engineering	2	1
1944	men	engineering	22	16
1945	men	engineering	25	19
1946	men	engineering	*	*
1947	men	engineering	35	25
1938	women	arts	19	14
1939	women	arts	16	11
1940	women	arts	18	15
1941	women	arts	21	15
1942	women	arts	9	8
1943	women	arts	13	13

1944	women	arts	22	18
1945	women	arts	22	18
1946	women	arts	1	1
1947	women	arts	16	13
1938	women	science	1	1
1939	women	science	4	4
1940	women	science	7	6
1941	women	science	3	3
1942	women	science	4	4
1943	women	science	9	8
1944	women	science	5	5
1945	women	science	17	16
1946	women	science	1	1
1947	women	science	10	10

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:

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

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

$$W_3 = \begin{cases} 1 & \text{if Faculty = science} \\ 0 & \text{otherwise} \end{cases}$$

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:

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

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
For this model, we have the following Minitab output:

```
Deviance Table
Source      DF  Adj Dev  Adj Mean  Chi-Square  P-Value
Regression   5   61.972    12.394       61.97    0.000
  year       1    3.445     3.445        3.45    0.063
  faculty    3   27.099     9.033       27.10    0.000
  sex        1   35.354    35.354       35.35    0.000
Error       52   54.114     1.041
Total       57  116.086

Goodness-of-Fit Tests
Test             DF  Chi-Square  P-Value
Deviance         52       54.11    0.394
Pearson          52       48.27    0.622
Hosmer-Lemeshow   7        9.89    0.195
```

The deviance of this model is:

$$D = 54.11 \text{ with } df = N - p = 58 - 6 = 52$$

(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:

$$H_o: \beta = 0 \text{ against } H_1: \beta \neq 0$$

The model under H_o is:

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

$p_o =$ Number of parameters $= 5$

For this model, we have the following Minitab output:

```
Goodness-of-Fit Tests
Test             DF  Chi-Square  P-Value
Deviance         53       57.56    0.310
Pearson          53       52.40    0.498
Hosmer-Lemeshow   4        0.73    0.947
```

The deviance of this model is:

$$D_o = 57.56 \text{ with } df = N - p_o = 58 - 5 = 53$$

Test statistic is:

$$\Delta D = D_o - D = 57.56 - 54.11 = 3.45 \text{ with } df = 53 - 52 = 1$$

Since $\Delta D = 3.45 < \chi^2_{0.05,(1)} = 3.84146$, we do not reject H_o 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?"

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 $= 5$

For this model, we have the following Minitab output:

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	4	25.435	6.359	25.44	0.000
year	1	2.176	2.176	2.18	0.140
faculty	3	20.436	6.812	20.44	0.000
Error	33	40.850	1.238		
Total	37	66.285			

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	33	40.85	0.164
Pearson	33	39.34	0.207
Hosmer-Lemeshow	7	10.53	0.161

The deviance of this model is:

$$D = 40.85 \text{ with } df = N - p = 38 - 5 = 33$$

To answer the question, we need to test:

$$H_o: \delta_1 = \delta_2 = \delta_3 = 0 \text{ against } H_1: \delta_j \neq 0 \text{ for at least one } \delta_j$$

The model under H_o is:

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

p_o = Number of parameters = 2

For this model, we have the following Minitab output:

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	36	61.29	0.005
Pearson	36	61.03	0.006
Hosmer-Lemeshow	4	2.84	0.584

The deviance of this model is:

$$D_o = 61.29 \text{ with } df = N - p_o = 38 - 2 = 36$$

Test statistic is:

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

Since $\Delta D = 20.44 > \chi^2_{0.05,(3)} = 7.81473$, we reject H_o 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) To answer the question " Are the proportions of female graduates who survived for 50 years after graduation the same for Arts and Science?"

We will use the data for women only.

Since there are only two faculties for women (Arts and Science) which means that the explanatory variable (Factor) "Faculty (W)" has 2 levels, therefore we define 1 dummy variable which is:

$$W = \begin{cases} 1; & \text{if Faculty} = \text{Science} \\ 0; & \text{if Faculty} = \text{Arts} \end{cases}$$

and 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

For this model, we have the following Minitab output:

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	2	10.605	5.3027	10.61	0.005

year	1	1.789	1.7889	1.79	0.181
faculty	1	7.063	7.0630	7.06	0.008
Error	17	11.950	0.7029		
Total	19	22.555			

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	17	11.95	0.803
Pearson	17	8.47	0.955
Hosmer-Lemeshow	7	4.15	0.762

The deviance of this model is:

$$D = 11.95 \text{ with } df = N - p = 20 - 3 = 17$$

To answer the question, we need to test:

$$H_o: \delta = 0 \text{ against } H_1: \delta \neq 0$$

The model under H_o is:

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

$p_o =$ Number of parameters $= 2$

For this model, we have the following Minitab output:

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	18	19.01	0.391
Pearson	18	13.89	0.736
Hosmer-Lemeshow	5	3.77	0.583

The deviance of this model is:

$$D_o = 19.01 \text{ with } df = N - p_o = 20 - 2 = 18$$

Test statistic is:

$$\Delta D = D_o - D = 19.01 - 11.95 = 7.06 \text{ with } df = 18 - 17 = 1$$

Since $\Delta D = 7.06 > \chi^2_{0.05,(1)} = 3.84146$, we reject H_o 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) To answer the question " 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?"

We will use the data for Arts and Science only.

Since there are only two faculties for women (Arts and Science), which means that the explanatory variable (Factor) "Faculty (W)" has 2 levels, therefore we define 1 dummy variable which is:

$$W = \begin{cases} 1; & \text{if Faculty} = \text{Science} \\ 0; & \text{if Faculty} = \text{Arts} \end{cases}$$

and 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 $= 5$.

For this model, we have the following Minitab output:

```
Deviance Table
Source         DF  Adj Dev  Adj Mean  Chi-Square  P-Value
Regression      4  53.8548   13.4637       53.85    0.000
  year          1   1.9417    1.9417        1.94    0.163
  sex           1  23.4687   23.4687       23.47    0.000
  faculty       1  17.0185   17.0185       17.02    0.000
  sex*faculty   1   0.8004    0.8004        0.80    0.371
Error          34  28.4163    0.8358
Total          38  82.2711

Goodness-of-Fit Tests
Test             DF  Chi-Square  P-Value
Deviance         34       28.42    0.738
Pearson          34       24.29    0.891
Hosmer-Lemeshow   7        2.67    0.913
```

The deviance of this model is:

$$D = 28.4163 \text{ with } df = N - p = 39 - 5 = 34$$

To answer the question, we need to test:

$$H_o: \tau = 0 \text{ against } H_1: \tau \neq 0$$

The model under H_o is:

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

$p_o =$ Number of parameters $= 4$.

For this model, we have the following Minitab output:

```
Goodness-of-Fit Tests
Test             DF  Chi-Square  P-Value
Deviance         35       29.22    0.743
Pearson          35       24.27    0.913
Hosmer-Lemeshow   7        3.26    0.860
```

The deviance of this model is:

$$D_o = 29.217 \text{ with } df = N - p_o = 39 - 4 = 35$$

Test statistic is:

$$\Delta D = D_o - D = 29.217 - 28.4163 = 0.8007 \text{ with } df = 35 - 34 = 1$$

Since $\Delta D = 0.8007 < \chi^2_{0.05,(1)} = 3.84146$, we do not reject H_o 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.

By using R:

```r
#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)
```

p= Number of parameters = 6

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

```r
#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
```

(a)

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
```

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

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

(b)

```r
#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
```

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

p= Number of parameters = 5

```r
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
```

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
```

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

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

(c)

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

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
```

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
```

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

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

(d)

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

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
```

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
```

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

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

Q 7.4:

(a) $D_0 - D_1 = 2[l(b_{\max}) - l(b_{\min})] - 2[l(b_{\max}) - l(b)] = C$

(b) For this hypothesis $D_0 \sim \chi^2(N-1), D_1 \sim \chi^2(N-p)$ so $C \sim \chi^2(p-1)$.