STAT 335: Generalized Linear Models
First Semester 1443

Chapter 7: Binary Variables and Logistic Regression: (Dobson 2008)

7.9 Exercises (p. 143)

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

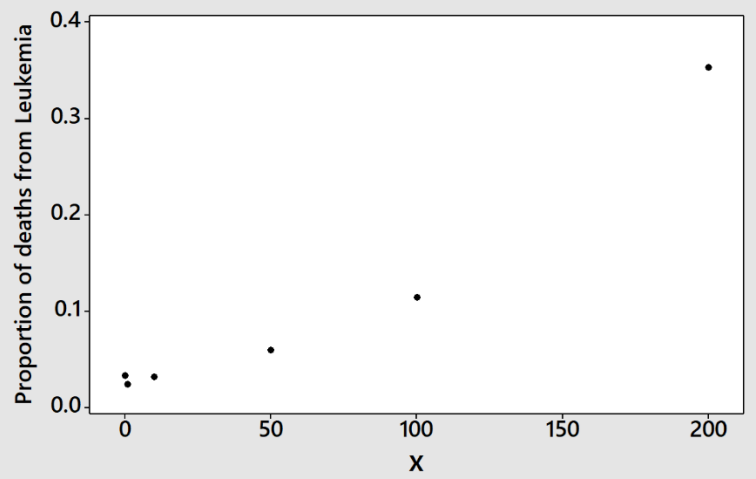
x_i = radiation dose (lower limit of radiation dose interval)

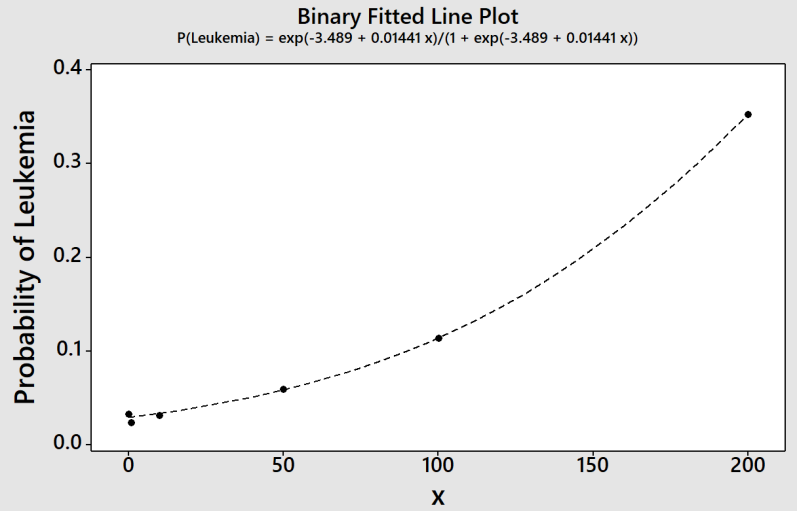
$i = 1, 2, \dots, N$

$N = 6$ (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:

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

- 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:

$$\hat{\beta}_1 = -3.489 \ , \ se(\hat{\beta}_1) = 0.204 \ , 95\% \ C.I \ for \ \beta_1 \ is \ (-3.889 \ , -3.089)$$
$$\hat{\beta}_2 = 0.01441 \ , \ se(\hat{\beta}_2) = 0.00182 \ , 95\% \ C.I \ for \ \beta_2 \ is \ (0.01085 \ , 0.01797)$$
$$Deviance: \ D = 0.4321 \ \ (with \ df = N - p = 6 - 2 = 4)$$
$$Pearson \ Chi - Square \ Statstics: \ X^2 = 0.4232 \ \ (with \ df = N - p = 6 - 2 = 4)$$

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

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

- The fitted equation of the model is:

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

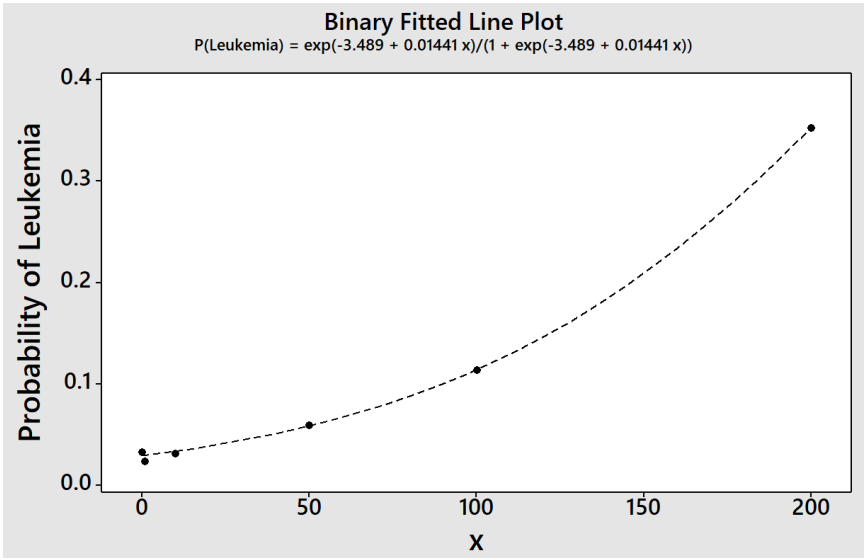
- The estimates of probabilities are:

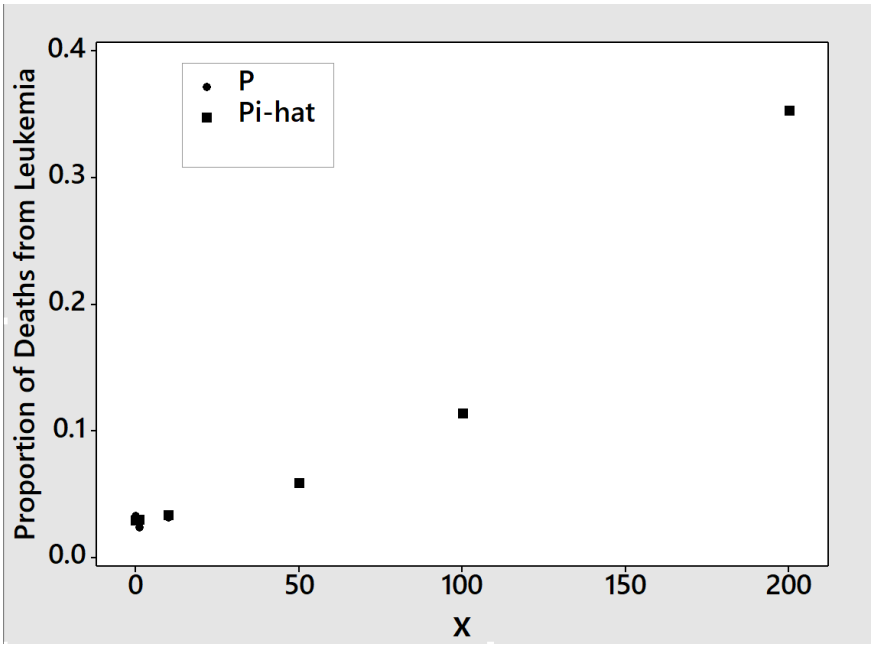
$$\hat{\pi}_i = \frac{e^{\widehat{\beta}_1 + \widehat{\beta}_2 x_i}}{1 + e^{\widehat{\beta}_1 + \widehat{\beta}_2 x_i}} = \frac{e^{-3.489 + 0.01441 x_i}}{1 + e^{-3.489 + 0.01441 x_i}}$$

- 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:

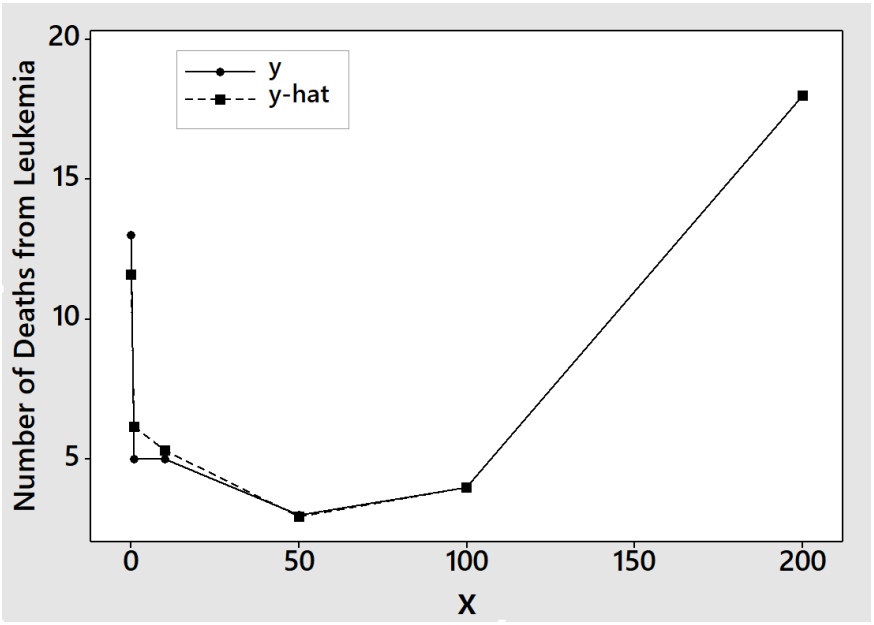
x_i	n_i	y_i	$P_i = \frac{y_i}{n_i}$	$\hat{\pi}_i$	e_{Di} Deviance Residual	e_{Pi} Pearson Residual	$\hat{y}_i$
0	391	13	0.0332	0.029628	0.4143	0.4222	11.584
1	205	5	0.0244	0.030045	-0.4899	-0.4743	6.159
10	156	5	0.0321	0.034064	-0.1399	-0.1386	5.314
50	50	3	0.0600	0.059052	0.0284	0.0284	2.953
100	35	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).



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	$1-\pi_1$	$O_1 = \frac{\pi_1}{1-\pi_1}$	$OR = \phi = \frac{O_1}{O_2}$
Not exposed	π_2	$1-\pi_2$	$O_2 = \frac{\pi_2}{1-\pi_2}$	

- The odds of disease for either exposure group is:

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

- 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 not exposed groups.

(a) For the simple logistic model is:

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

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}	$1-\pi_{1j}$	$O_{1j} = \frac{\pi_{1j}}{1-\pi_{1j}}$	$\phi_j = \frac{O_{1j}}{O_{2j}}$
Not exposed	π_{2j}	$1-\pi_{2j}$	$O_{2j} = \frac{\pi_2}{1-\pi_{2j}}$	

Consider the following logistic model:

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

$$\Leftrightarrow$$

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

$$\Leftrightarrow$$

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

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:

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

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:

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:

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

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

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

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:

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

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

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:

$$D = 54.11 \quad with\ df = N - p = 58 - 6 = 52$$

(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:

$$H_o: \beta = 0 \quad against \quad H_1: \beta \neq 0$$

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:

$$D_o = 57.56 \quad with\ df = N - p_o = 58 - 5 = 53$$

Test statistic is:

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

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:

$$D = 40.85 \quad with\ df = N - p = 38 - 5 = 33$$

To answer the question, we need to test:

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

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:

$$D_o = 61.29 \quad with\ df = N - p_o = 38 - 2 = 36$$

Test statistic is:

$$\Delta D = D_o - D = 61.29 - 40.85 = 20.44 \quad with \quad df = 36 - 33 = 3$$

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:

$$D = 11.95 \quad with\ df = N - p = 20 - 3 = 17$$

To answer the question, we need to test:

$$H_o: \delta = 0 \quad against \quad H_1: \delta \neq 0$$

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:

$$D_o = 19.01 \quad with\ df = N - p_o = 20 - 2 = 18$$

Test statistic is:

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

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:

$$D = 28.4163 \quad with\ df = N - p = 39 - 5 = 34$$

To answer the question, we need to test:

$$H_o: \tau = 0 \quad against \quad H_1: \tau \neq 0$$

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:

$$D_o = 29.217 \quad with\ df = N - p_o = 39 - 4 = 35$$

Test statistic is:

$$\Delta D = D_o - D = 29.217 - 28.4163 = 0.8007 \quad with \quad df = 35 - 34 = 1$$

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.