Chapter Seven

Ryan Honea

10/9/2017

Exercise One

Question

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

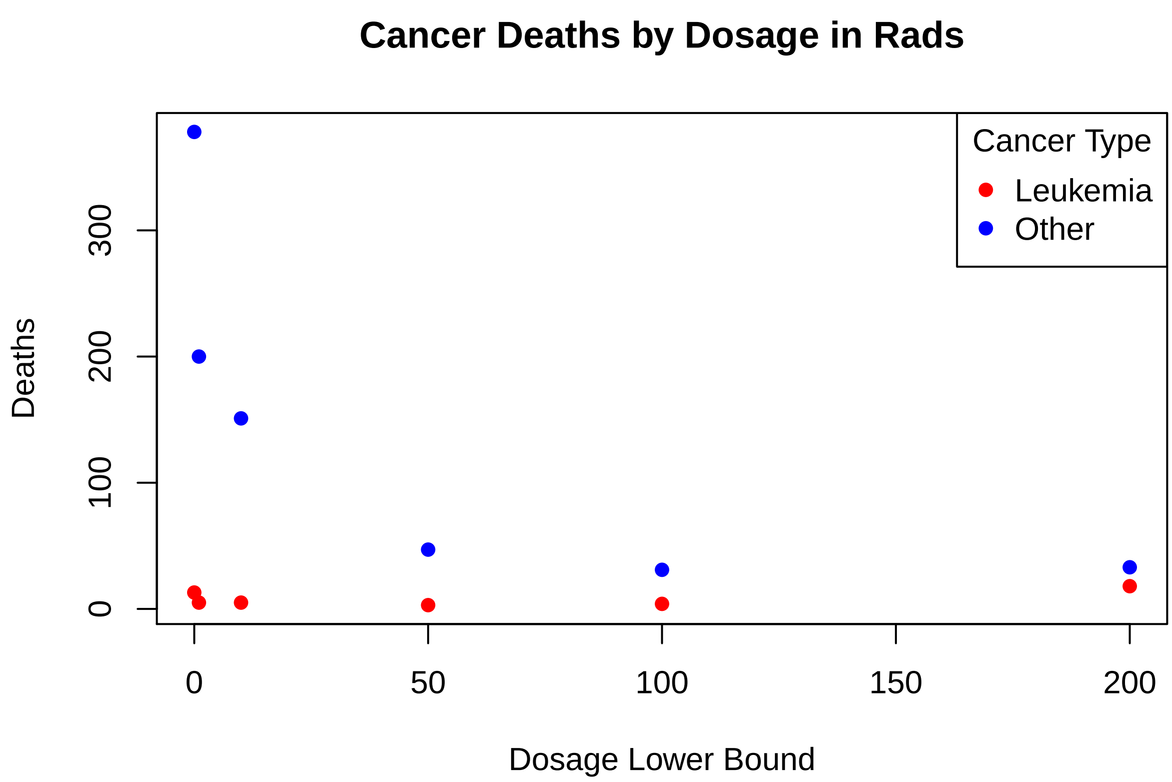
	Radiation Dose (rads)					
Deaths	0	1-9	10-49	50-99	100-199	200+
Leukemia	13	5	5	3	4	18
Other Cancers	378	200	151	47	31	33
Total Cancers	391	205	156	50	35	51

Solution

(a): Obtain a suitable model to describe the dose-response relationship between radiation and the proportional cancer mortality rates for leukemia.

Solution: First we enter the data and plot it to gain an initial understanding of our data.

```
Dosage <- c("0", "1-9", "10-49","50-99","100-199","200+")
Dose.Bounded <- c(0,1,10,50,100,200)
Leukemia <- c(13, 5, 5, 3, 4, 18)
Other <- c(378, 200, 151, 47, 31, 33)
labels <- c("Leukemia","Other")
colors <- c(rep("red",6), rep("blue",6))
plot(c(Dose.Bounded, Dose.Bounded), c(Leukemia, Other),
     pch = c(16,16), col = colors,
     xlab = "Dosage Lower Bound", ylab = "Deaths",
     main = "Cancer Deaths by Dosage in Rads")
legend("topright", title = "Cancer Type",
       labels, col = c("red","blue"), pch = c(16,16))
```



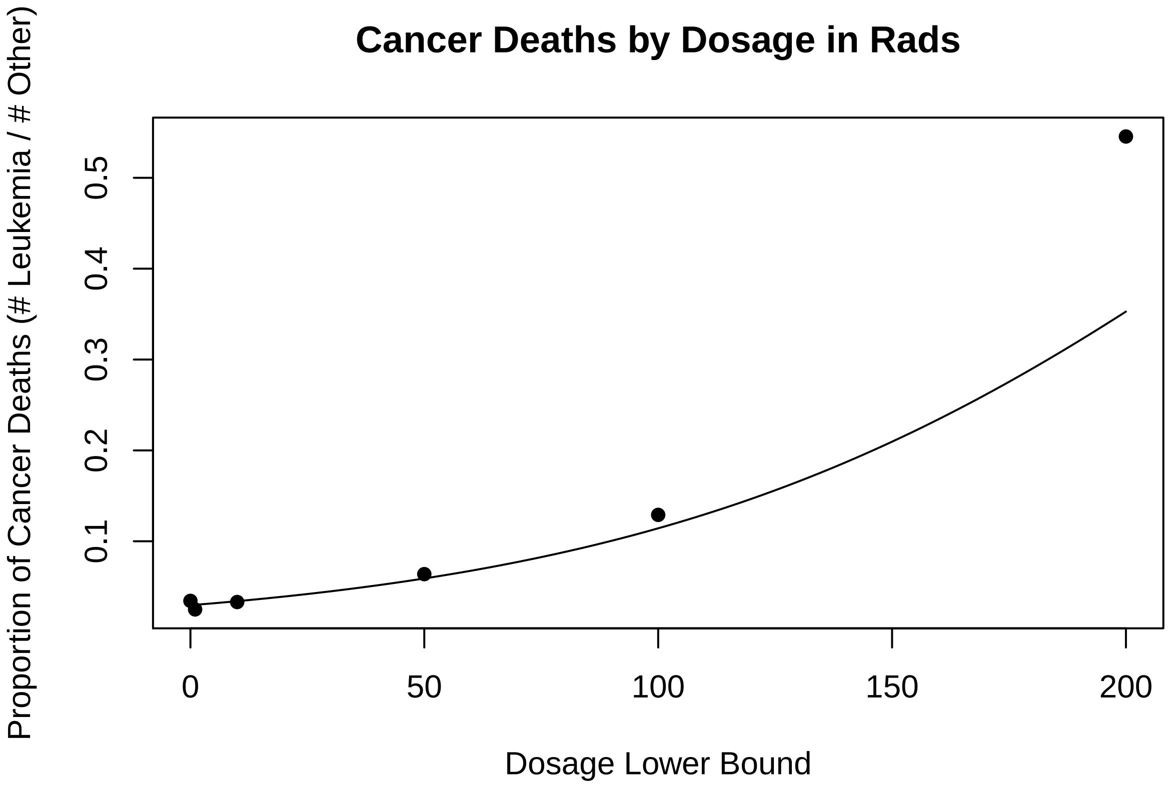
There is clearly a divide here that we can hope to find with a logistic model.

```
cancers <- cbind(Leukemia, Other)
logmod <- glm(cancers ~ Dose.Bounded, family=binomial(link = "logit"))
logmod
```

```
##
## Call:  glm(formula = cancers ~ Dose.Bounded, family = binomial(link = "logit"))
##
## Coefficients:
##  (Intercept)  Dose.Bounded
##     -3.48897       0.01441
##
## Degrees of Freedom: 5 Total (i.e. Null);  4 Residual
## Null Deviance:       54.35
## Residual Deviance: 0.4321    AIC: 26.1
```

(b): Examine how well the model describes the data.

Solution: Below is a graph checking the accuracy of the model by drawing a line along the proportion of cancer deaths.



As there are six covariate patterns and one parameter, we can find the Deviance and check it against $\chi^2(6-1)$.

```
deviance = logmod$deviance
p.value <- 1 - pchisq(deviance, df = 5)
cat("P.Value:",p.value)
```

```
## P.Value: 0.9944005
```

Conversely, one can use the Hosmer Lemeshow test where $D \sim \chi^2(g-2)$ where in this case, we have six groups, so $g = 6$.

```
deviance = logmod$deviance
p.value <- 1 - pchisq(deviance, df = 4)
cat("P.Value:",p.value)
```

```
## P.Value: 0.9797692
```

(c): Interpret the results.

Solution: Based on the above p-values, it is strongly suggested that this is a strong model with a good fit.

Exercise Two

Question

Odds ratios. Consider a 2×2 contingency table from a prospective study in which people who were or were not exposed to some pollutants are followed up and, after several years, categorized according to the presence or absence of a disease. Table 7.13 (below) shows the probabilities for each cell. The odds of disease for either exposure group is $O_i = \pi_i/(1-\pi_i)$, for $i = 1, 2$, and so the odds ratio

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

is a measure of the relative likelihood of disease for the exposed and not exposed groups.

	Diseased	Not diseased
Exposed	π_1	$1-\pi_1$
Not exposed	π_2	$1-\pi_2$

Solution

(a): For the simple logistic model $\pi_i = e^{\beta_i}/(1+e^{\beta_i})$, show that if there is no difference between the exposed and not exposed groups (i.e. $\beta_1 = \beta_2$), then $\phi = 1$.

Solution: If $\beta_1 = \beta_2$, then

$$\pi_1 = \frac{e^{\beta_1}}{1+e^{\beta_1}} = \frac{e^{\beta_2}}{1+e^{\beta_2}} = \pi_2$$

and thus

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

(b): Consider J 2×2 tables like Table 7.13, one for each level x_j of a factor, such as age group, with $j = 1, ..., J$. For the logistic model

$$\pi_{ij} = \frac{\exp(\alpha_i + \beta_i x_j)}{1+\exp(\alpha_i + \beta_i x_j)}, \quad i = 1, 2, \quad j = 1, ..., J.$$

Show that $\log \phi$ is constant over all table if $\beta_1 = \beta_2$ (McKinlay 1979).

Solution: If $\beta_1 = \beta_2$ and $\eta_i = \alpha_i + \beta_i x_j$

$$\begin{aligned}
\phi_j = \frac{\pi_{1j}(1-\pi_{2j})}{\pi_{2j}(1-\pi_{1j})} &= \frac{\frac{\exp(\eta_1)}{1+\exp(\eta_1)}\left(1-\frac{\exp(\eta_2)}{1+\exp(\eta_2)}\right)}{\frac{\exp(\eta_2)}{1+\exp(\eta_2)}\left(1-\frac{\exp(\eta_1)}{1+\exp(\eta_1)}\right)} \\
&= \frac{\frac{\exp(\eta_1)}{1+\exp(\eta_1)}\left(\frac{1}{1+\exp(\eta_2)}\right)}{\frac{\exp(\eta_2)}{1+\exp(\eta_2)}\left(\frac{1}{1+\exp(\eta_1)}\right)} = \frac{\eta_1}{\eta_2} \\
&= \exp(\eta_1 - \eta_2) = \exp(\alpha_1 - \alpha_2 - x_j(\beta_1 - \beta_2)) \\
&= \exp(\alpha_1 - \alpha_2 - x_j(\beta_1 - \beta_1)) = \exp(\alpha_1 - \alpha_2)
\end{aligned}$$

So we have shown that ϕ_j is constant when $\beta_1 = \beta_2$.

Exercise Three

Question

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

Year of Graduation	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	103	142

Year of Graduation	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

Solution

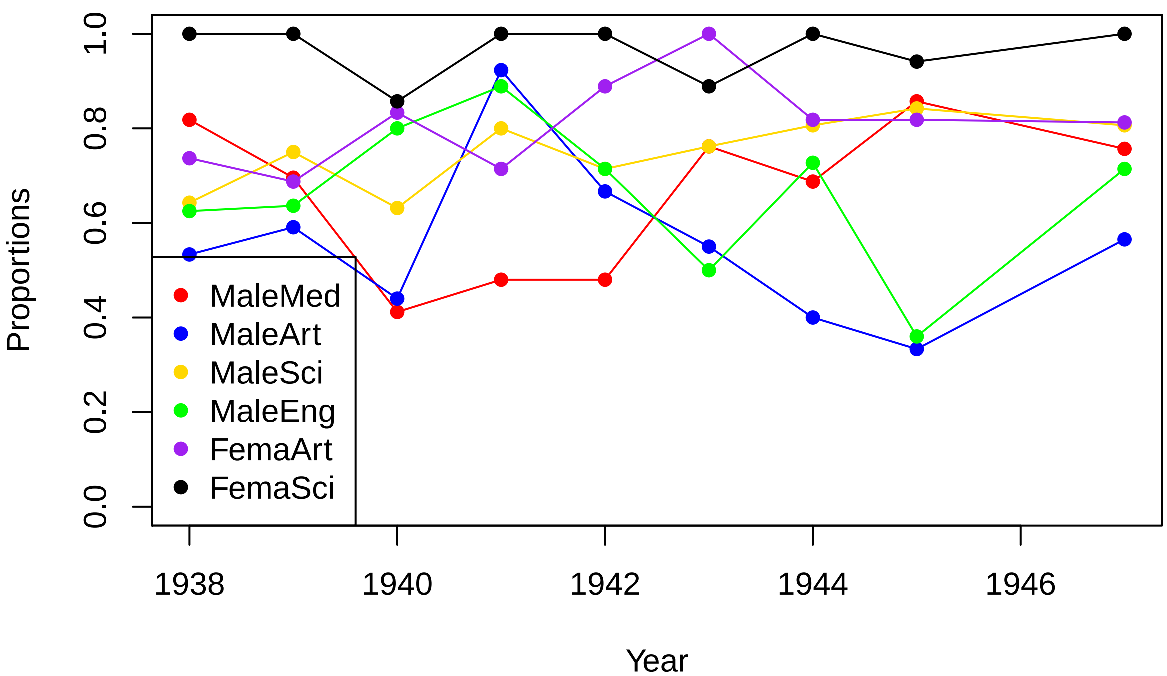
Data Entry:

```
YoG <- seq(1938,1947,by=1)
Male.Med.S <- c(18,16,7,12,24,16,22,12,22,28)
Male.Med.T <- c(22,23,17,25,50,21,32,14,34,37)
Male.Art.S <- c(16,13,11,12,8,11,4,4,0,13)
Male.Art.T <- c(30,22,25,13,12,20,10,12,0,23)
Male.Sci.S <- c(9,9,12,12,20,16,25,32,4,25)
Male.Sci.T <- c(14,12,19,15,28,21,31,38,5,31)
Male.Eng.S <- c(10,7,12,8,5,1,16,9,0,25)
Male.Eng.T <- c(16,11,15,9,7,2,22,25,0,35)
Fema.Art.S <- c(14,11,15,15,8,13,18,18,1,13)
```

```
Fema.Art.T <- c(19,16,18,21,9,13,22,22,1,16)
Fema.Sci.S <- c(1,4,6,3,4,8,5,16,1,10)
Fema.Sci.T <- c(1,4,7,3,4,9,5,17,1,10)
df <- data.frame(YoG = YoG, Male.Med.S = Male.Med.S, Male.Med.T = Male.Med.T,
                 Male.Art.S = Male.Art.S, Male.Art.T = Male.Art.T,
                 Male.Sci.S = Male.Sci.S, Male.Sci.T = Male.Sci.T,
                 Male.Eng.S = Male.Eng.S, Male.Eng.T = Male.Eng.T,
                 Fema.Art.S = Fema.Art.S, Fema.Art.T = Fema.Art.T,
                 Fema.Sci.S = Fema.Sci.S, Fema.Sci.T = Fema.Sci.T)
```

Because the year 1946 is missing values, I'm just going to remove the row.

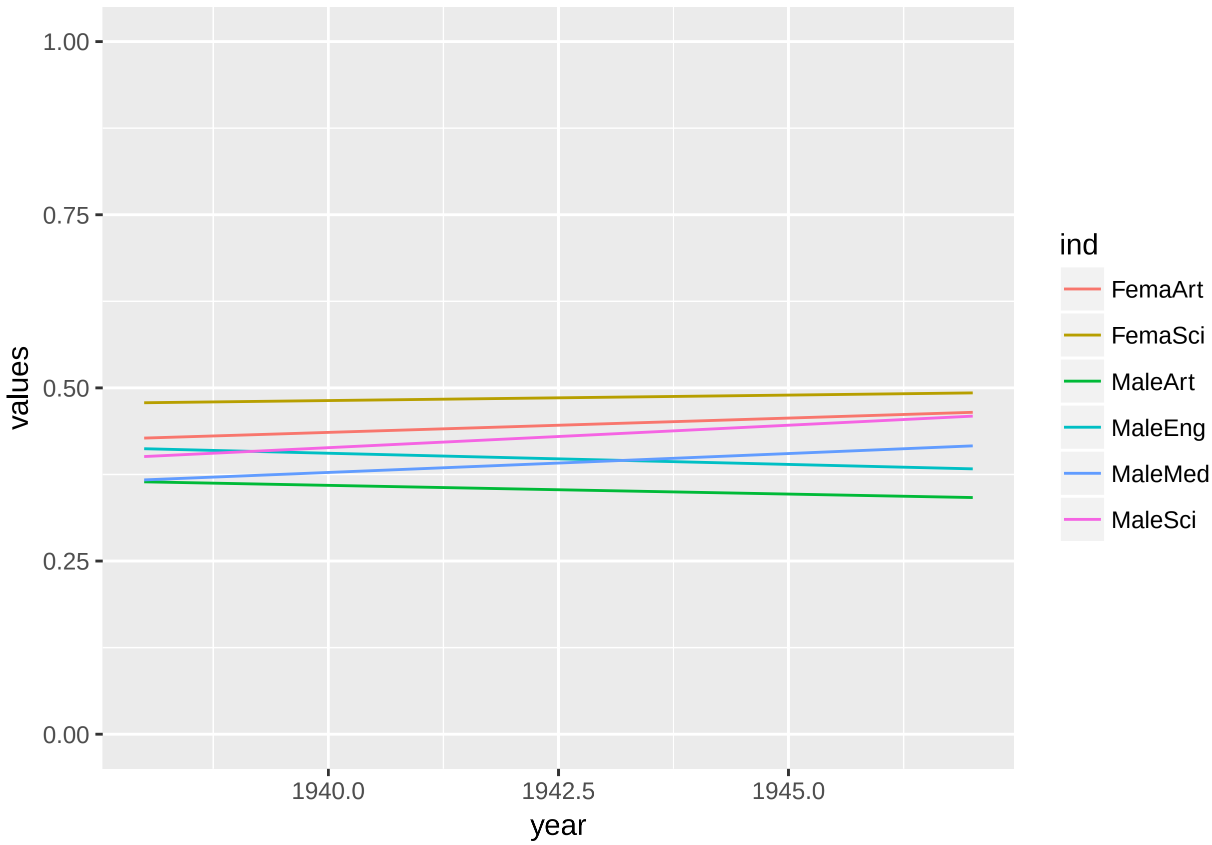
```
df <- df[-9,]
proportions <- matrix(, nrow = 9, ncol = 6)
for (i in 1:6) {
  proportions[1:9,i] = df[,2*i]/df[,2*i + 1]
}
labels = c("MaleMed", "MaleArt", "MaleSci", "MaleEng", "FemaArt", "FemaSci")
colors = c("red"    , "blue"   , "gold"   , "green"  , "purple" , "black")
plot(df[,1], proportions[,1], col = colors[1],
     ylim = c(0,1), pch = 16, type = "o",
     xlab = "Year", ylab = "Proportions")
for (i in 2:6) {
  lines(df[,1], proportions[,i], col = colors[i], pch = 16, type = "o")
}
legend("bottomleft", labels, col = colors, pch = rep(16,6))
```



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

```
models <- list()
for (i in 1:6) {
  models[[i]] <- glm(cbind(df[,i*2],df[,i*2 + 1]) ~ df[,1], family = binomial(link = "logit"))
}
```

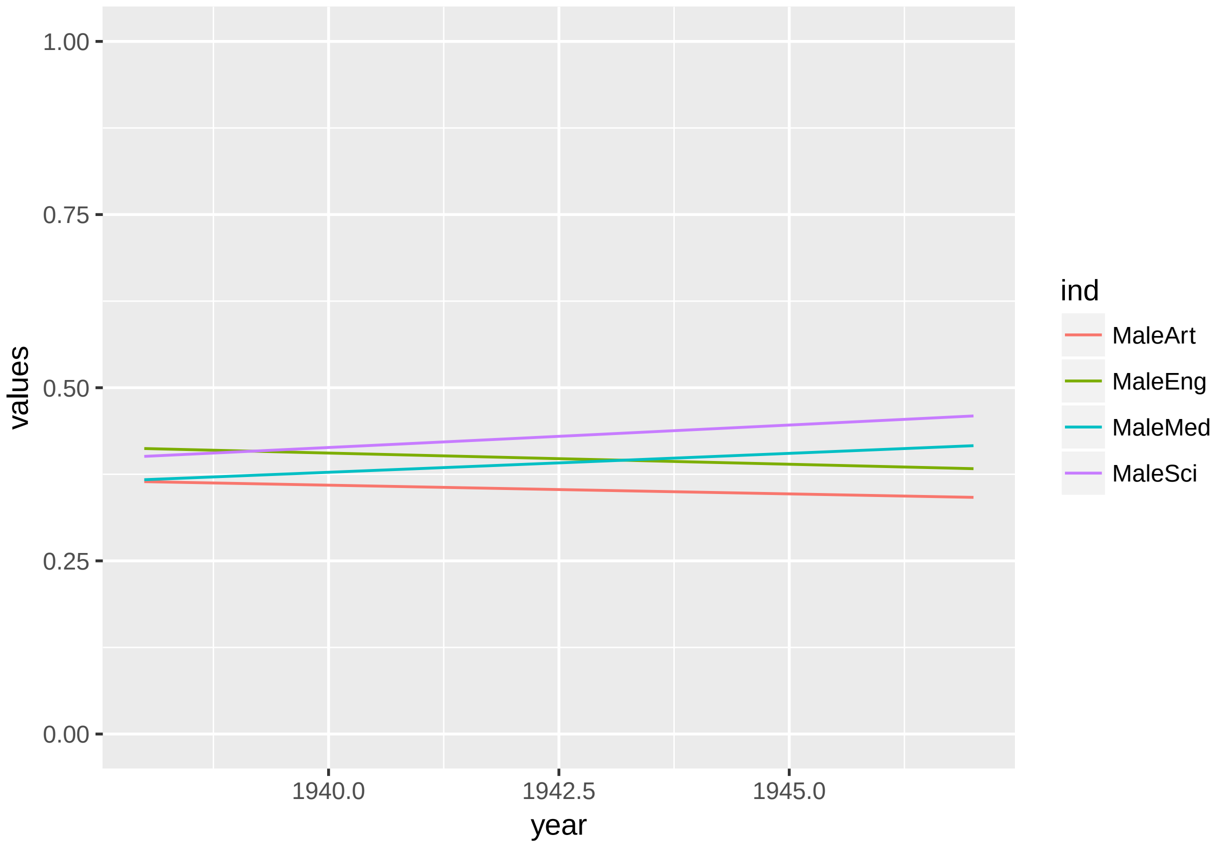
```
predval <- sapply(models,
    function(m) predict(m, data.frame(x=df[,1]), type="response"))
  predval <- data.frame(df[,1], predval)
  colnames(predval) <- c("Year", labels)
obs <- stack(predval[,2:7])
year <- df[,1]
obs <- cbind(year, obs)
ggplot(obs, aes(x=year, y=values, color=ind))+geom_line()+ylim(0,1)
```



I would argue that none of these visually seem to be increasing much or decreasing much so I would argue that the proportions are equal across the years.

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

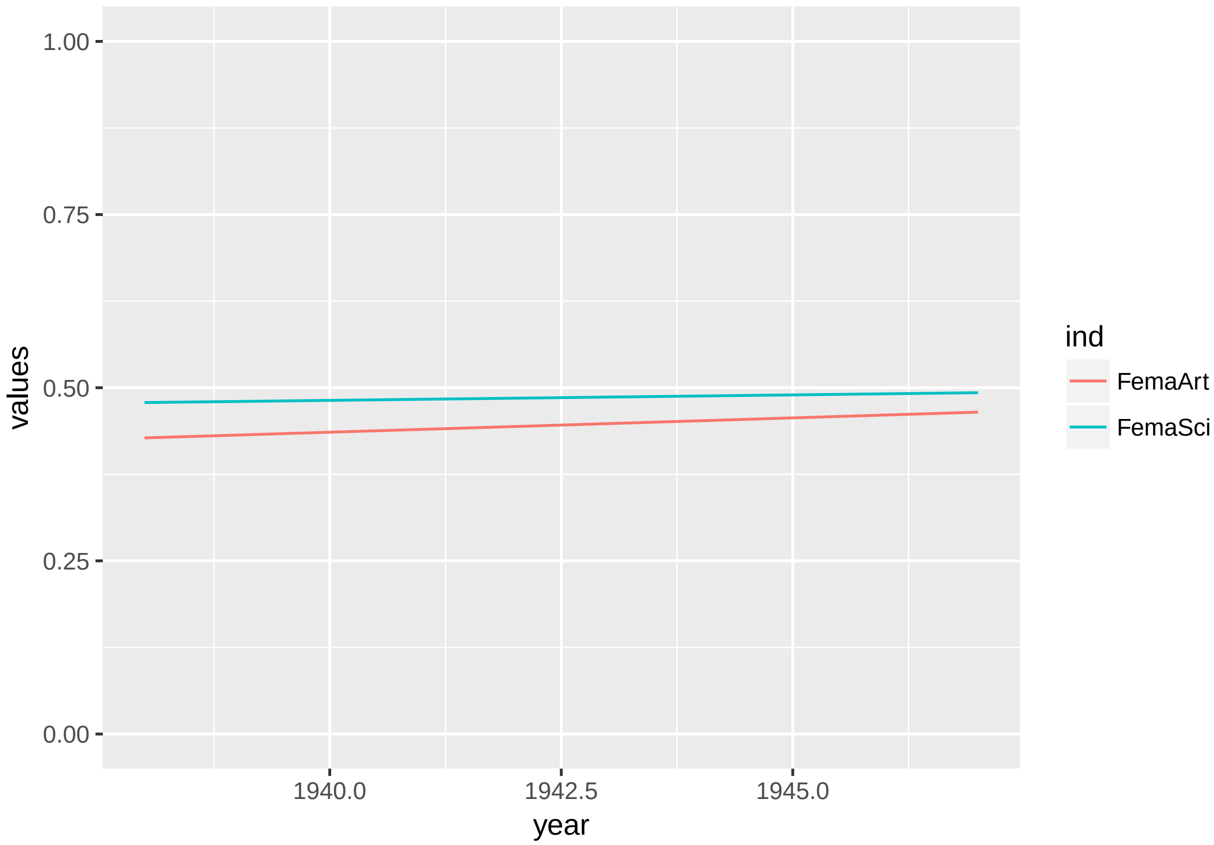
```
predval <- sapply(models,
    function(m) predict(m, data.frame(x=df[,1]), type="response"))
  predval <- data.frame(df[,1], predval)
  colnames(predval) <- c("Year", labels)
obs <- stack(predval[,2:5])
year <- df[,1]
obs <- cbind(year, obs)
ggplot(obs, aes(x=year, y=values, color=ind))+geom_line()+ylim(0,1)
```



Males in the sciences appear to live longer than the art counterparts over the years, but the only conclusive evidence I offer is this graph.

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

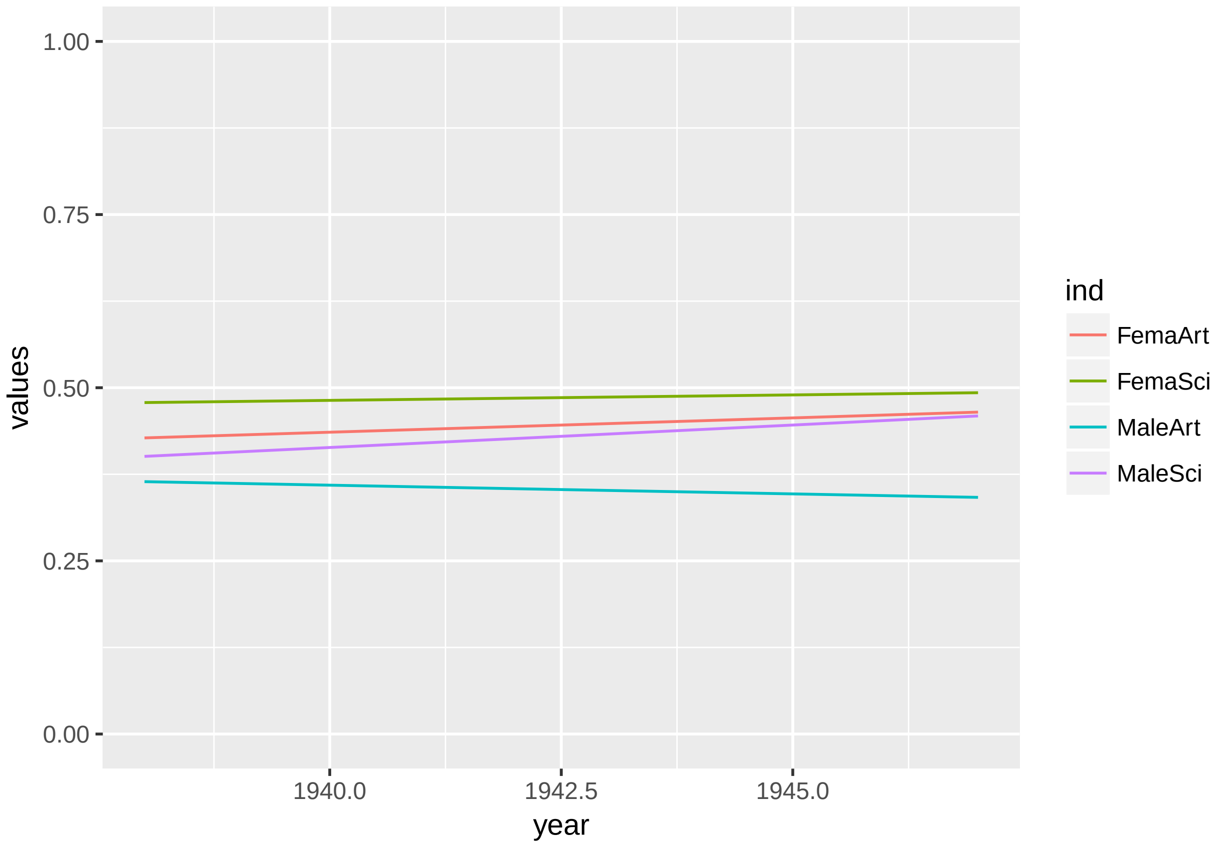
```
predval <- sapply(models,
    function(m) predict(m, data.frame(x=df[,1]), type="response"))
  predval <- data.frame(df[,1], predval)
  colnames(predval) <- c("Year", labels)
obs <- stack(predval[,6:7])
year <- df[,1]
obs <- cbind(year, obs)
ggplot(obs, aes(x=year, y=values, color=ind))+geom_line()+ylim(0,1)
```



The same is here. Women in the sciences appear to live longer than women in the arts.

(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?

```
predval <- sapply(models,
    function(m) predict(m, data.frame(x=df[,1]), type="response"))
  predval <- data.frame(df[,1], predval)
  colnames(predval) <- c("Year", labels)
obs <- stack(c(predval[,3:4],predval[,6:7]))
year <- df[,1]
obs <- cbind(year, obs)
ggplot(obs, aes(x=year, y=values, color=ind))+geom_line()+ylim(0,1)
```



The differences appear to be much wider with the men, but again the only conclusive evidence I offer is the graph above.