MATH3823 Assessed Practical - Outline Analysis

The following is not a draft report, but is a *log* of a typical R session. Your report will then have summarized the main ideas, with the R commands moved to an appendix.

Initial data analysis

As instructed, read in the data with a command such as:

```
adelaide = read.csv("http://rgaykroyd.github.io/MATH3823/Datasets/adelaide-00.csv")
attach(adelaide)
```

though 00 should be replaced by the appropriate last two digits of the Student Identification Number (SID).

The following assumed the SID=42.

It is always a good idea to inspect the data, for example using:

```
str(adelaide)
```

```
## 'data.frame':    60 obs. of  5 variables:
##  $ year   : int  1938 1939 1940 1941 1942 1943 1944 1945 1946 1947 ...
##  $ faculty: chr  "M" "M" "M" "M" ...
##  $ sex    : chr  "M" "M" "M" "M" ...
##  $ survive: int  9 9 8 24 29 16 16 14 24 21 ...
##  $ total  : int  12 19 18 39 45 23 29 19 35 29 ...
```

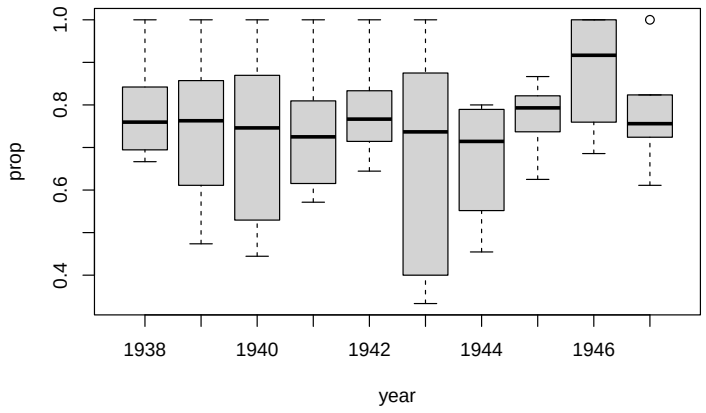
The data set contains 60 rows of information. However, a look at the data shows that the value of `total` is 0 in two rows of the table. For statistical fitting purposes, R ignores these rows. Hence the effective sample size is 58, not 60. Thus, in all the statistical models fitted to the whole data set, the degrees of freedom of the NULL model is always 58 - 1 = 57, not 60 - 1 = 59.

By default, in R, the variable variable `year` is numeric (that is quantitative rather than a factor). Conversely the variables `faculty` and `sex` are factors, because their entries include letters rather than numbers. In cases were letters or words are used, rather than numbers, there is no need ro declare variables as factors – but it does no harm.

Any analysis should start with a few plots. The two exploratory plots below can be constructed to investigate patterns in the data.

```
# Define the proportion surviving and plot split by year and then by faculty
prop=survive/total

boxplot(prop~year,main="Fig 1: Boxplots of proportion surviving by year")
```

Fig 1: Boxplots of proportion surviving by year




```
tmp = lm(prop~year)
itrend = tmp$coefficients[2]
```

Fig 1 shows that the proportion surviving 50 years does not change much from year to year. Perhaps there is a slight increasing trend – this can be tested later.

```
boxplot(prop~faculty+sex,main="Fig 2: Boxplots of prop. surviving by faculty and sex")
```

Fig 2: Boxplots of prop. surviving by faculty and sex

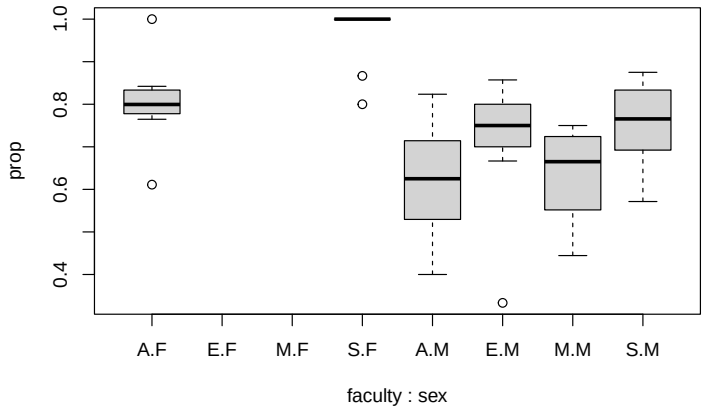


Fig 2 shows that: (a) in general, the proportion of females surviving seems to be greater than for males, (b) no females did Engineering or Medicine, and (c) for males, Engineers and Science graduates seem to survive better than Arts and Medicine graduates.

From this preliminary exploration, there is clearly variation in the proportion surviving and, further, that sex and faculty seem to be the most important variables.

The initial statistical model

From the problem description it is clear that the appropriate model has `survive` as the response variable (though `total` is also important). Each year there are a fixed number of graduates, variable `total`, who are

studied and after the 50-year period each graduate is either still alive or has died. Further, it is reasonable to assume that graduates behave independently. This is a classical setting for a binomial model where the *number of trials* is the number of graduates each year and we expect the probability of survival to depend on `sex`, `faculty` and `year` as the explanatory variables.

The question sheet asks for the following model, Model 1 say, to be fitted with corresponding output:

```
died=total-survive
ym=cbind(survive,died)

M1 = glm(ym ~ year+faculty+sex,family=binomial)

summary(M1)
```

```
##
## Call:
## glm(formula = ym ~ year + faculty + sex, family = binomial)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -45.41370   48.74618  -0.932 0.351525
## year          0.02412    0.02510   0.961 0.336670
## facultyE      0.56731    0.25963   2.185 0.028884 *
## facultyM     -0.02019    0.20201  -0.100 0.920398
## facultyS      0.65002    0.20654   3.147 0.001649 **
## sexM         -0.87092    0.22609  -3.852 0.000117 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 86.664  on 57  degrees of freedom
## Residual deviance: 50.377  on 52  degrees of freedom
## AIC: 211.05
##
## Number of Fisher Scoring iterations: 4
```

Notice that it is necessary to treat the data as binomial responses, with `ym` being a two column matrix containing the number of survivors and non-survivors. It is also possible to consider other link functions, but for simplicity we stick to the standard logit link function here.

[Some students fitted a Gaussian regression with `prop` as the response variable. This is wrong - see the discussion of the Beetle data set in lecture notes.]

The column "estimates" contains the information needed to calculate the systematic linear predictor, η, of the model for any combination of explanatory variables. The intercept is always present; the coefficient for year states how much the linear predictor increases for each unit increase in year, the coefficients for faculty indicate the contribution for each faculty (with Arts having a value 0); similarly the coefficient for sexM indicates the contribution for Male (with Female having a value 0).

The coefficient for year has a p-value (p=0.3367) bigger than 0.05. Hence, we can say that, although there is a hint of increasing probability of survival with year, but it is not significant. Below we consider a simpler model, removing year.

The fitted model can now be used for prediction. For example, with explanatory variables `sex=M`, `year=1941`, `faculty=M`:

```
predict(M1, newdata = data.frame(sex="M", year=1941, faculty="M"), type="response")
```

```
##         1
## 0.6237243
```

and

```
predict(M1, newdata = data.frame(sex="F", year=1938, faculty="E"), type="response")
```

```
##         1
## 0.8689205
```

Note the latter probability is larger because: (a) females live longer than males (because the estimate for `sexM` is negative), and (b) Engineers live longer than doctors (because the estimate for `facultyE` is larger than for `facultyM`). Note that these two effects more than compensates for the opposite effect for year: (c) students in 1941 live a bit longer than students in 1938 (because the estimate for `year` is positive).

Recall that no women studied Engineering. Hence the second fitted probability is an *extrapolated* probability of survival; that is, it is the fitted probability of surviving 50 years if a woman had done Engineering, assuming the fitted model is true.

Further statistical analysis

A good starting point is to judge the goodness of fit of the model just fitted by looking at the residual deviance, $D_1 = 50.38 \sim \chi^2$ with $df = 52$. The p-value is 0.5379 (which is greater than 0.05), and hence the model fits the data well.

Since it appears that `year` is not important, it is reasonable to consider a model with this removed, Model 2, with results:

```
M2=glm( ym ~ faculty+sex, family=binomial) # remove year

summary(M2)
```

```
##
## Call:
## glm(formula = ym ~ faculty + sex, family = binomial)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.421067   0.188187   7.551 4.31e-14 ***
## facultyE     0.595160   0.257890   2.308 0.021010 *
## facultyM     0.005267   0.200066   0.026 0.978998
## facultyS     0.678908   0.204227   3.324 0.000886 ***
## sexM        -0.875503   0.225893  -3.876 0.000106 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 86.664  on 57  degrees of freedom
```

```
## Residual deviance: 51.301  on 53  degrees of freedom
## AIC: 209.97
##
## Number of Fisher Scoring iterations: 4
```

Comparing the deviance of Model 2 to Model 1, the change in deviance is $D_2 - D_1 = 51.3\text{-}50.38 = 0.92 \sim \chi^2$ ($df = 1$) with p-value 0.3363 (which is greater than 0.05). Hence, we accept the hypothesis that the simpler Model 2 is adequate, and that the more complicated Model 1 is not significantly better.

From now, there is no single correct analysis but you were expected to try other sensible alternatives and write about them in a logical way. What follows are some alternative models which you might have included.

Another potential model is to include the interaction between `faculty` and `sex` (with `year` excluded), Model 3 say:

```
M3=glm( ym ~ faculty*sex, family=binomial) # remove year

summary(M3)
```

```
##
## Call:
## glm(formula = ym ~ faculty * sex, family = binomial)
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.28262    0.19674   6.519 7.06e-11 ***
## facultyE       0.50264    0.26431   1.902  0.05721 .
## facultyM      -0.08726    0.20827  -0.419  0.67525
## facultyS       1.64412    0.62442   2.633  0.00846 **
## sexM          -0.64453    0.25690  -2.509  0.01211 *
## facultyE:sexM       NA         NA      NA       NA
## facultyM:sexM       NA         NA      NA       NA
## facultyS:sexM -1.14205    0.66422  -1.719  0.08554 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 86.664  on 57  degrees of freedom
## Residual deviance: 47.698  on 52  degrees of freedom
## AIC: 208.37
##
## Number of Fisher Scoring iterations: 5
```

Comparing the deviance of Model 3 to Model 2, the change in deviance is $D_2 - D_3 = 51.3\text{-}47.7 = 3.6 \sim \chi^2$ ($df = 1$) with p-value 0.0576 (which is greater than 0.05). Hence, we accept the hypothesis that the simpler Model 2 is adequate and the more complicated Model 3 is not significantly better.

It is now reasonable to consider the removal, in turn, of `sex` and `faculty` to see if an even simpler model is adequate.

First consider the model with `faculty` only:

```
M4 = glm( ym ~ faculty, family=binomial)

deviance(M4)-deviance(M2)
```

```
## [1] 16.1743
```

```
df.residual(M4)-df.residual(M2)
```

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

Comparing the deviance of Model 4 to Model 2, the change in deviance is $D_4 - D_2 = 67.48\text{-}51.3 = 16.17 \sim \chi^2$ ($df = 1$) with p-value 0.0001 (which is less than 0.05). Hence, we reject the hypothesis that the simpler Model 4 is adequate and hence conclude that variable `sex` is important.

Next consider the model with `sex` only:

```
M5=glm( ym ~ sex,     family=binomial)
deviance(M5)-deviance(M2)
```

```
## [1] 19.03702
```

```
df.residual(M5)-df.residual(M2)
```

```
## [1] 3
```

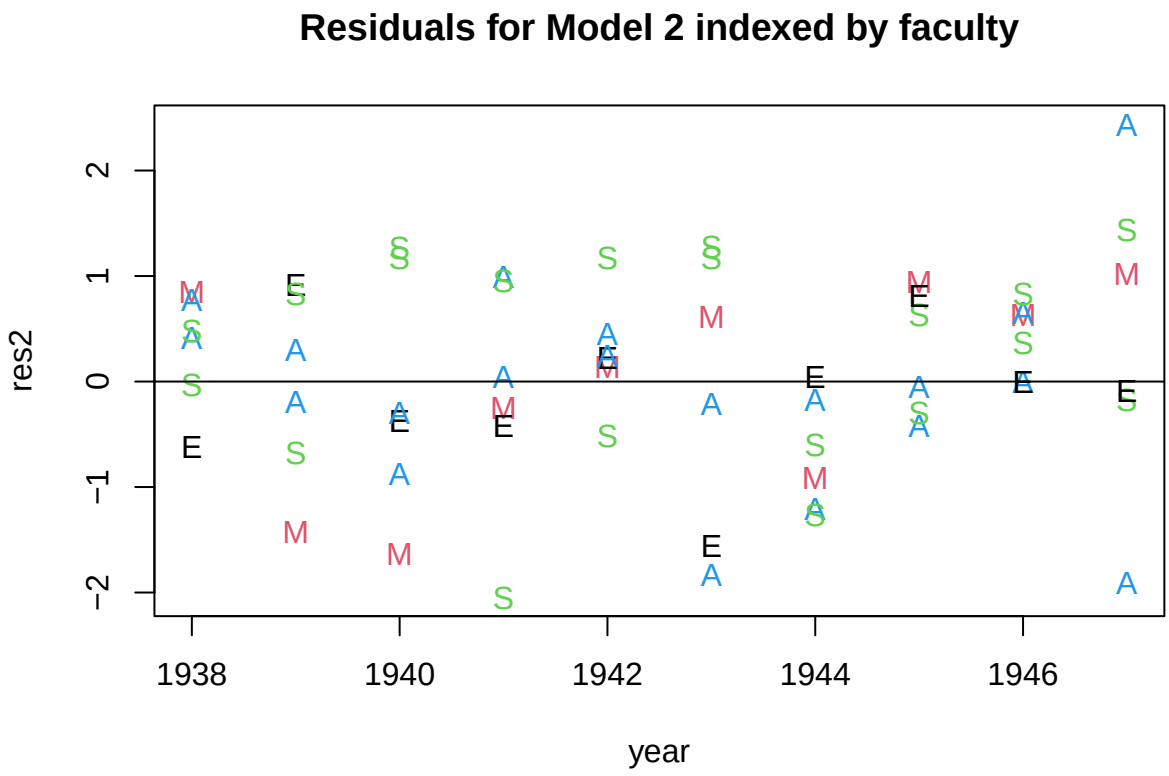
Comparing the deviance of Model 5 to Model 2, the change in deviance is $D_5 - D_2 = 70.34\text{-}51.3 = 19.04 \sim \chi^2$ ($df = 3$) with p-value 0.0003 (which is less than 0.05). Hence, we reject the hypothesis that the simpler Model 5 is adequate and hence that `faculty` is important.

Before moving to further models, consider the residuals for the best fit model so far considered.

```
res2=residuals(M2)

plot(year,res2, pch=as.character(sex), col=1+(sex=="F"),
     main="Residuals for Model 2 indexed by sex")

abline(h=0)
```

The plot of deviance residuals against year, with symbols distinguishing between Male and Female graduates, shows no clear pattern. Similarly, a residual plot with plotting symbols distinguishing faculty shows no pattern.

```
plot(year,res2, pch=as.character(faculty),
     col=1+(faculty=="M")+2*(faculty=="S")+3*(faculty=="A"),
     main="Residuals for Model 2 indexed by faculty")

abline(h=0)
```



In conclusion, together the testing results and residual plots confirm that the best model so far considered has response `survive` modelled as a binomial with probability of survival depending on `sex` and `faculty` which can be written as:

$$\texttt{survive}_{ijk} \sim \text{Bin}(m, p_{ij}) \text{ with } \text{logit}(p_{ij}) = \alpha + \beta_i + \gamma_j$$

where $i = \{A, E, M, S\}$, $j = \{F, M\}$ and $k = \{1938, \ldots, 1947\}$.

Then, from the output for Model 2 which is duplicated below, we have: $\hat{\alpha} = 1.42$, $\hat{\boldsymbol{\beta}} = \{\ 0,\ 0.6,\ 0.01,\ 0.68\ \}$ and $\hat{\gamma} = \{\ 0,\ \text{-}0.88\ \}$.

```
summary(M2)
```

```
##
## Call:
## glm(formula = ym ~ faculty + sex, family = binomial)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.421067   0.188187   7.551 4.31e-14 ***
## facultyE     0.595160   0.257890   2.308 0.021010 *
## facultyM     0.005267   0.200066   0.026 0.978998
## facultyS     0.678908   0.204227   3.324 0.000886 ***
## sexM        -0.875503   0.225893  -3.876 0.000106 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 86.664  on 57  degrees of freedom
## Residual deviance: 51.301  on 53  degrees of freedom
## AIC: 209.97
##
## Number of Fisher Scoring iterations: 4
```

Examining these results further, note that the parameter `facultyM` is not significant. That is not significantly different to `facultyA` which is fixed at zero and hence we might merge the catgories.

```
faculty.red = faculty
faculty.red[faculty=="A"] = "AM"
faculty.red[faculty=="M"] = "AM"

M6 = glm(formula = ym ~ faculty.red + sex, family = binomial)

summary(M6)
```

```
##
## Call:
## glm(formula = ym ~ faculty.red + sex, family = binomial)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.4216     0.1870   7.601 2.94e-14 ***
## faculty.redE   0.5920     0.2284   2.591  0.00956 **
## faculty.redS   0.6761     0.1738   3.890  0.00010 ***
## sexM          -0.8729     0.2031  -4.298 1.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 86.664  on 57  degrees of freedom
## Residual deviance: 51.302  on 54  degrees of freedom
## AIC: 207.97
##
## Number of Fisher Scoring iterations: 4
```

```
deviance(M6)-deviance(M2)
```

```
## [1] 0.0006929937
```

```
df.residual(M6)-df.residual(M2)
```

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

The change in deviance is $D_6 - D_2 = 0 \sim \chi^2$ ($df = 1$) with p-value 0.979 (which is greater than 0.05). Hence, we accept the null hypothesis and conclude that the simpler model (with levels combined) is adequate.

Also that `facultyS` and `facultyE` have similar parameter estimates and hence we can combining.

```
faculty.red2 = faculty.red
faculty.red2[faculty.red=="E"] = "ES"
faculty.red2[faculty.red=="S"] = "ES"

M7 = glm(formula = ym ~ faculty.red2 + sex, family = binomial)

summary(M7)
```

```
##
## Call:
## glm(formula = ym ~ faculty.red2 + sex, family = binomial)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.4270     0.1863   7.659 1.88e-14 ***
## faculty.red2ES   0.6486     0.1518   4.274 1.92e-05 ***
## sexM            -0.8796     0.2019  -4.355 1.33e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 86.664  on 57  degrees of freedom
## Residual deviance: 51.410  on 55  degrees of freedom
## AIC: 206.08
##
## Number of Fisher Scoring iterations: 4
```

```
deviance(M7)-deviance(M2)
```

```
## [1] 0.1086926
```

```
df.residual(M7)-df.residual(M2)
```

```
## [1] 2
```

```
pchisq(1.350863, 2, lower.tail = F)
```

```
## [1] 0.5089368
```

The change in deviance is $D_7 - D_2 = 0.11 \sim \chi^2$ ($df = 2$) with p-value 0.9471 (which is greater than 0.05). Hence, again,we accept the null hypothesis and conclude that the simpler model (with further levels combined) is adequate. This has lead us to a model whee faculty is only included as either "*Arts/Medicine*" or "*Engineering/Science*" – which does seem to make some sense.

Although there was little pattern seen in survival from year to year, the following model considers `year` again, but this time as a factor which can be a good alternative if the relationship is non-linear.

```
yearf=as.factor(year)

M7 = glm(ym ~ yearf+faculty+sex,family=binomial) # year as factor
summary(M7)
```

```
##
## Call:
## glm(formula = ym ~ yearf + faculty + sex, family = binomial)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.6492164  0.2998125   5.501 3.78e-08 ***
## yearf1939   -0.3413528  0.3456627  -0.988   0.3234
## yearf1940   -0.3826337  0.3377164  -1.133   0.2572
## yearf1941   -0.3459751  0.3276125  -1.056   0.2909
## yearf1942   -0.1540665  0.3422440  -0.450   0.6526
## yearf1943   -0.2357806  0.3567124  -0.661   0.5086
## yearf1944   -0.5548742  0.3259367  -1.702   0.0887 .
## yearf1945   -0.0074059  0.3479841  -0.021   0.9830
## yearf1946    0.0920536  0.4402189   0.209   0.8344
## yearf1947    0.0005035  0.3142430   0.002   0.9987
## facultyE     0.6260230  0.2660289   2.353   0.0186 *
## facultyM     0.0140661  0.2139098   0.066   0.9476
## facultyS     0.6902235  0.2115538   3.263   0.0011 **
## sexM        -0.9122525  0.2300416  -3.966 7.32e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 86.664  on 57  degrees of freedom
## Residual deviance: 43.600  on 44  degrees of freedom
## AIC: 220.27
##
## Number of Fisher Scoring iterations: 4
```

```
M8 = glm(ym ~ faculty+sex,family=binomial) # remove year
summary(M8)
```

```
##
## Call:
## glm(formula = ym ~ faculty + sex, family = binomial)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.421067   0.188187   7.551 4.31e-14 ***
## facultyE     0.595160   0.257890   2.308 0.021010 *
## facultyM     0.005267   0.200066   0.026 0.978998
## facultyS     0.678908   0.204227   3.324 0.000886 ***
## sexM        -0.875503   0.225893  -3.876 0.000106 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 86.664  on 57  degrees of freedom
## Residual deviance: 51.301  on 53  degrees of freedom
## AIC: 209.97
##
```

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

```
deviance(M8)-deviance(M7)
```

```
## [1] 7.700998
```

```
df.residual(M8)-df.residual(M7)
```

```
## [1] 9
```

The difference between the deviances is $D_8 - D_7 = 7.7 \sim \chi^2$ ($df = 9$) with p-value 0.5645 (which is greater than 0.05). Hence, the simpler model without `year` is an adequate fit and treating `year` as a factor has not altered this conclusion. There is little point in considering models with years combined unless, for example, we suspected, or saw from the data, clear periods were the survival was constant with occasional abrupt changes.

Seperate analysis of Male and Female graduates

In this part we split the data into males and females and fit models separately.

```
#########################################################
# analyze two sexes separately
totalm=total[sex=="M"]; survivem=survive[sex=="M"]; diedm=totalm-survivem
ymm=cbind(survivem,diedm); yearm=year[sex=="M"]; facultym=faculty[sex=="M"]

totalf=total[sex=="F"]; survivef=survive[sex=="F"]; diedf=totalf-survivef
ymf=cbind(survivef,diedf); yearf=year[sex=="F"]; facultyf=faculty[sex=="F"]

M9m=glm(ymm ~ facultym, family=binomial)
M9f=glm(ymf ~ facultyf, family=binomial)

summary(M9m)
```

```
##
## Call:
## glm(formula = ymm ~ facultym, family = binomial)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.63809    0.16520   3.863 0.000112 ***
## facultymE    0.50264    0.26431   1.902 0.057209 .
## facultymM   -0.08726    0.20827  -0.419 0.675251
## facultymS    0.50207    0.22647   2.217 0.026624 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 49.153  on 37  degrees of freedom
## Residual deviance: 36.531  on 34  degrees of freedom
## AIC: 163.08
##
## Number of Fisher Scoring iterations: 4
```

```r
summary(M9f)
```

```
##
## Call:
## glm(formula = ymf ~ facultyf, family = binomial)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.2826     0.1967   6.519 7.06e-11 ***
## facultyfS     1.6441     0.6244   2.633  0.00846 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21.185  on 19  degrees of freedom
## Residual deviance: 11.167  on 18  degrees of freedom
## AIC: 45.285
##
## Number of Fisher Scoring iterations: 5
```

The first thing to notice is that the qualitative interpretation is similar to Model 2. For Males: Engineering and Science graduates live longest; Medicine and Arts graduates live shortest. Further, it looks as if combining level of `faculty` as earlier would also be sensible.

For Females, of course, there are no estimates for Engineers and Medical graduates, but there is no significant difference between Arts and Science graduates.

```r
M10f=glm(ymf ~ 1, family=binomial)

deviance(M10f)-deviance(M9f)
```

```
## [1] 10.01853
```

```r
df.residual(M10f)-df.residual(M9f)
```

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

Fitting the simpler model without `faculty` gives a change in deviance of $10.02 \sim \chi^2$ ($df = 1$) with p-value 0.0015 (which is less than 0.05). Hence, the simpler model is not adequate and faculty is important for the survival of females.

So, the conclusions have not changed. In particular, the previous overall conclusion about survival of graduates from different faculties still holds.

End of MATH3823 Assessed Practical - Outline Analysis