Practice 8 Solutions

1. The following data on marital status by gender and report of premarital sex (PMS) and extramarital sex (EMS) were obtained by the UK Marriage Research Centre.

			Marital Status	
Gender	PMS	EMS	Divorced	Still Married
Male	Yes	Yes	28	11
		No	60	42
	No	Yes	17	4
		No	68	130
Females	Yes	Yes	17	4
		No	54	25
	No	Yes	36	4
		No	214	322

The data are stored in file `Mstatus.csv`.

(a) Fit a saturated log-linear model to the data and see what terms are significant.

```
> marr=read.csv("data/Mstatus.csv")
> marr
```

```
   freq gend PMS EMS MS freq.b gend.b PMS.b EMS.b N.b
1    28    M   Y   Y  D     28      M     Y     Y  39
2    60    M   Y   N  D     60      M     Y     N 102
3    17    M   N   Y  D     17      M     N     Y  21
4    68    M   N   N  D     68      M     N     N 198
5    17    F   Y   Y  D     17      F     Y     Y  21
6    54    F   Y   N  D     54      F     Y     N  79
7    36    F   N   Y  D     36      F     N     Y  40
8   214    F   N   N  D    214      F     N     N 536
9    11    M   Y   Y  M     NA                    NA
10   42    M   Y   N  M     NA                    NA
11    4    M   N   Y  M     NA                    NA
12  130    M   N   N  M     NA                    NA
13    4    F   Y   Y  M     NA                    NA
14   25    F   Y   N  M     NA                    NA
15    4    F   N   Y  M     NA                    NA
16  322    F   N   N  M     NA                    NA
```

```
> ms1=glm(freq~gend*PMS*EMS*MS, family=poisson, data=marr)
> anova(ms1,test="Chi")
```

```
Analysis of Deviance Table
Model: poisson, link: log
Response: freq

Terms added sequentially (first to last)
```

```
               Df   Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                                15    1333.85
gend            1      97.94        14    1235.92  4.314e-23
PMS             1     312.29        13     923.63  6.933e-70
EMS             1     689.26        12     234.36 6.463e-152
MS              1       2.22        11     232.14       0.14
gend:PMS        1      75.26        10     156.88  4.128e-18
gend:EMS        1      12.76         9     144.12  3.537e-04
PMS:EMS         1      36.16         8     107.96  1.816e-09
gend:MS         1       0.03         7     107.93       0.86
PMS:MS          1      45.46         6      62.47  1.558e-11
EMS:MS          1      48.84         5      13.63  2.783e-12
gend:PMS:EMS    1  1.144e-03         4      13.63       0.97
gend:PMS:MS     1       0.63         3      13.00       0.43
gend:EMS:MS     1       2.67         2      10.33       0.10
PMS:EMS:MS      1      10.19         1       0.15  1.415e-03
gend:PMS:EMS:MS 1       0.15         0  2.531e-14       0.70
```

(b) Use the `step` procedure to find the "best" model. Use `scopems=list(lower=freq ~ gend*PMS*EMS, upper=freq ~ gend*PMS*EMS*MS)` to define the range of the models to be examined. Note that we may treat `gend`, `PMS` and `EMS` as the "explanatory" factors and `MS` as the response factor here. So we want to always keep the 3-factor interaction `gend:PMS:EMS` term in the model.

```
> scopems=list(lower=freq~gend*PMS*EMS, upper=freq~gend*PMS*EMS*MS)
> step.ms2=step(ms1, scope=scopems)

Step:  AIC=110.7
freq ~ gend + PMS + EMS + MS + gend:PMS + gend:EMS + PMS:EMS +
    gend:MS + PMS:MS + EMS:MS + gend:PMS:EMS + PMS:EMS:MS

              Df Deviance     AIC
<none>             0.698 110.702
+ gend:EMS:MS  1    0.290 112.294
+ gend:PMS:MS  1    0.440 112.444
- gend:MS      1    5.246 113.250
- PMS:EMS:MS   1   13.629 121.633

> anova(step.ms2,test="Chi")

Analysis of Deviance Table
Model: poisson, link: log
Response: freq

Terms added sequentially (first to last)

             Df   Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                            15    1333.85
gend          1      97.94        14    1235.92  4.314e-23
PMS           1     312.29        13     923.63  6.933e-70
EMS           1     689.26        12     234.36 6.463e-152
MS            1       2.22        11     232.14       0.14
gend:PMS      1      75.26        10     156.88  4.128e-18
gend:EMS      1      12.76         9     144.12  3.537e-04
PMS:EMS       1      36.16         8     107.96  1.816e-09
```

```
gend:MS       1       0.03         7     107.93       0.86
PMS:MS        1      45.46         6      62.47  1.558e-11
EMS:MS        1      48.84         5      13.63  2.783e-12
gend:PMS:EMS  1  1.144e-03         4      13.63       0.97
PMS:EMS:MS    1      12.93         3       0.70  3.231e-04
```

(c) Based on the "best" model found in (b), test whether the effect of `gend:MS` is significant.

```
> ms2=glm(freq~gend*PMS*EMS+PMS*EMS*MS+gend*MS, family=poisson, data=marr)
> ms3=glm(freq~gend*PMS*EMS+PMS*EMS*MS, family=poisson, data=marr)

> anova(ms3,ms2,test="Chi")

Analysis of Deviance Table

Model 1: freq ~ gend * PMS * EMS + PMS * EMS * MS
Model 2: freq ~ gend * PMS * EMS + PMS * EMS * MS + gend * MS
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         4     5.2455
2         3     0.6978  1   4.5477    0.0330
```

- The `gend:MS` interaction is significant. The "best" model suggests that marital status `MS` has associations with all of the explanatory factors `gend`, `PMS` and `EMS`. Among these associations, `PMS` and `EMS` interact in their association with `MS`, while the association between `gend` and `MS` is not influenced by `PMS` or `EMS`.

(d) Regard `MS` as the response factor, fit logistic regression models to the data, find the "best" logistic model and explain it.

- First fit a saturated logistic model to the data. (Note the results are equivalent to those from the saturated log-liner model).

```
> marr2=marr[1:8, 6:10]
> marr2

  freq.b gend.b PMS.b EMS.b N.b
1     28      M     Y     Y  39
2     60      M     Y     N 102
3     17      M     N     Y  21
4     68      M     N     N 198
5     17      F     Y     Y  21
6     54      F     Y     N  79
7     36      F     N     Y  40
8    214      F     N     N 536
```

```
> msb1=glm(freq.b/N.b~gend.b*PMS.b*EMS.b, family=binomial, weight=N.b, data=marr2)
> anova(msb1, test="Chi")

Analysis of Deviance Table
Model: binomial, link: logit
Response: freq.b/N.b

Terms added sequentially (first to last)

                  Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
```

```
NULL                                    7    107.956
gend.b              1    0.031          6    107.925      0.861
PMS.b               1   45.460          5     62.465  1.558e-11
EMS.b               1   48.836          4     13.629  2.783e-12
gend.b:PMS.b        1    0.630          3     13.000      0.428
gend.b:EMS.b        1    2.667          2     10.332      0.102
PMS.b:EMS.b         1   10.186          1      0.146      0.001
gend.b:PMS.b:EMS.b  1    0.146          0 -9.326e-15      0.702
```

```
> summary(msb1)$coef
```

```
                        Estimate Std. Error    z value     Pr(>|z|)
(Intercept)           -0.4085755 0.08819573 -4.6325998 3.611022e-06
gend.bM               -0.2394512 0.17371436 -1.3784193 1.680739e-01
PMS.bY                 1.1786838 0.25748205  4.5777317 4.700449e-06
EMS.bY                 2.6058001 0.53437322  4.8763673 1.080573e-06
gend.bM:PMS.bY        -0.1739821 0.35940421 -0.4840847 6.283258e-01
gend.bM:EMS.bY        -0.5108544 0.78535097 -0.6504791 5.153828e-01
PMS.bY:EMS.bY         -1.9289893 0.80802009 -2.3873037 1.697246e-02
gend.bM:PMS.bY:EMS.bY  0.4116779 1.07294801  0.3836886 7.012093e-01
```

```
> summary(ms1)$coef
```

```
                        Estimate Std. Error      z value     Pr(>|z|)
(Intercept)          5.365976015 0.06835859 78.497461736 0.000000e+00
gendM               -1.146468310 0.13920768 -8.235668241 1.785581e-16
PMSY                -1.376991968 0.15228728 -9.042068166 1.537352e-19
EMSY                -1.782457077 0.18014071 -9.894804391 4.384667e-23
MSM                  0.408575531 0.08819573  4.632599767 3.611022e-06
gendM:PMSY           1.251828826 0.23358931  5.359101566 8.363683e-08
gendM:EMSY           0.396162715 0.32554583  1.216918419 2.236353e-01
PMSY:EMSY            0.626686374 0.33134985  1.891313289 5.858254e-02
gendM:MSM            0.239451215 0.17371436  1.378419268 1.680739e-01
PMSY:MSM            -1.178683752 0.25748205 -4.577731748 4.700449e-06
EMSY:MSM            -2.605800108 0.53437464 -4.876354384 1.080644e-06
gendM:PMSY:EMSY     -0.002532065 0.48549262 -0.005215455 9.958387e-01
gendM:PMSY:MSM       0.173982063 0.35940421  0.484084655 6.283258e-01
gendM:EMSY:MSM       0.510854380 0.78535213  0.650478127 5.153834e-01
PMSY:EMSY:MSM        1.928989347 0.80802122  2.387300360 1.697262e-02
gendM:PMSY:EMSY:MSM -0.411677912 1.07294900 -0.383688237 7.012095e-01
```

- The three interaction terms involving `gend` in model `msb1` may be non-significant. Now omit these three terms and compare the resultant model with the saturated model.

```
> msb2=glm(freq.b/N.b~gend.b+PMS.b*EMS.b, family=binomial, weight=N.b, data=marr2)
> anova(msb2,msb1,test="Chi")
```

```
Analysis of Deviance Table

Model 1: freq.b/N.b ~ gend.b + PMS.b * EMS.b
Model 2: freq.b/N.b ~ gend.b * PMS.b * EMS.b
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         3    0.69784
2         0 -9.326e-15  3  0.69784   0.87371
```

- None of the three omitted interaction terms are significant. It seems model `msb2` is the "best" model. The procedure `step(msb1)` also verifies this. Note

the equivalent results from `anova(msb2)` and `anova(step.ms2)`, and `summary(msb2)$coef` and `sumamry(ms2)$coef`.

```
> step(msb1)

Step:  AIC=45.96
freq.b/N.b ~ gend.b + PMS.b + EMS.b + PMS.b:EMS.b

              Df Deviance    AIC
<none>             0.698 45.961
- gend.b       1    5.246 48.509
- PMS.b:EMS.b  1   13.629 56.893

Call:  glm(formula = freq.b/N.b ~ gend.b + PMS.b + EMS.b + PMS.b:EMS.b,
           family = binomial, data = marr2, weights = N.b)

Coefficients:
  (Intercept)        gend.bM         PMS.bY         EMS.bY  PMS.bY:EMS.bY
      -0.3908        -0.3089         1.0995         2.3960        -1.7999

Degrees of Freedom: 7 Total (i.e. Null);  3 Residual
Null Deviance:      108
Residual Deviance: 0.6978       AIC: 45.96

> anova(msb2,test="Chi")

Analysis of Deviance Table

Model: binomial, link: logit

Response: freq.b/N.b

Terms added sequentially (first to last)

            Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                            7    107.956
gend.b       1    0.031         6    107.925     0.861
PMS.b        1   45.460         5     62.465 1.558e-11
EMS.b        1   48.836         4     13.629 2.783e-12
PMS.b:EMS.b  1   12.931         3      0.698 3.231e-04

> summary(msb2)$coef

                Estimate Std. Error   z value     Pr(>|z|)
(Intercept)   -0.3908113 0.08458134 -4.620538 3.827467e-06
gend.bM       -0.3088840 0.14582031 -2.118251 3.415381e-02
PMS.bY         1.0994964 0.17867453  6.153627 7.573074e-10
EMS.bY         2.3960414 0.38789243  6.177077 6.529920e-10
PMS.bY:EMS.bY -1.7998744 0.51295439 -3.508839 4.500673e-04

> summary(ms2)$coef

                  Estimate Std. Error     z value     Pr(>|z|)
(Intercept)      5.3766099 0.06641510  80.9546302 0.000000e+00
gendM           -1.1913254 0.12620531  -9.4395823 3.742683e-21
PMSY            -1.4074747 0.13860165 -10.1548191 3.153947e-24
EMSY            -1.8140365 0.17659492 -10.2723025 9.393551e-25
MSM              0.3908113 0.08458134   4.6205378 3.827467e-06
```

```
gendM:PMSY       1.3340678 0.17697940   7.5379837 4.772935e-14
gendM:EMSY       0.5049230 0.29201380   1.7291068 8.378998e-02
PMSY:EMSY        0.6494524 0.31298163   2.0750495 3.798194e-02
PMSY:MSM        -1.0994964 0.17867453  -6.1536269 7.573074e-10
EMSY:MSM        -2.3960414 0.38789243  -6.1770771 6.529919e-10
gendM:MSM        0.3088840 0.14582031   2.1182511 3.415381e-02
gendM:PMSY:EMSY -0.1030026 0.42304425  -0.2434796 8.076339e-01
PMSY:EMSY:MSM    1.7998744 0.51295439   3.5088390 4.500673e-04
```

Based on the "best" model `msb2`,

- odds of a male being divorced is $e^{-0.309} = 0.734$ times that of a female;
- compared to `PMS`=no and `EMS`=no:
  i. odds of `PMS`=yes and `EMS`=no being divorced is $e^{1.0995} = 3.00$ times as large,
  ii. odds of `PMS`=no and `EMS`=yes being divorced is $e^{2.396} = 10.98$ times as large,
  iii. odds of `PMS`=yes, `EMS`=yes being divorced is $e^{1.0995+2.396-1.7999} = 5.45$ times as large.

2. Carry out an analysis of the following 1959 data (`mobilityDenmark.csv`) on the occupational mobility of males in Denmark, and state your conclusions. (Hint: a similar analysis for the UK mobility data can be found in the lecture notes.)

Status Category of Father's Occupation	Status Category of Sons's Occupation 1	2	3	4	5	Total
1	18	17	16	4	2	57
2	24	105	109	59	21	318
3	23	84	289	217	95	708
4	8	49	175	348	198	778
5	6	8	69	201	246	530
Total	79	263	658	829	562	2391

- Start with the no association model, then add a term (`diag`) for the main diagonal, and then one for the distance (the factor `dist`) from the main diagonal. Since the analysis of deviance table (below) shows the contributions of the terms added sequentially we assess the adequacy of three models (no association; association down the main diagonal only; association as a distance from the main diagonal (as a factor)) by fitting a single model.

```
> mobDen=read.csv("D:/data/mobilityDenmark.csv")
> mob1=glm(freq~factor(F)+factor(S)+diag+factor(dist),family=poisson,data=mobDen)
> anova(mob1, test="Chi")

Analysis of Deviance Table
Terms added sequentially (first to last)

             Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
```

```
NULL                          24    2489.01
factor(F)     4   920.03      20    1568.98 7.614e-198
factor(S)     4   914.77      16     654.21 1.049e-196
diag          1   303.61      15     350.60  5.381e-68
factor(dist)  3   338.18      12      12.41  5.402e-73

> 1-pchisq(12.41,12)

[1] 0.4133387
```

- The model with distance as a factor provides a good fit to the data (the `diag` term is redundant as it is implied by `dist`). See if distance as a factor can be replaced by distance as a variable.

```
> mob2=glm(freq~factor(F)+factor(S)+dist+factor(dist),family=poisson,data=mobDen)
> anova(mob2, test="Chi")

Analysis of Deviance Table

             Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                          24    2489.01
factor(F)     4   920.03      20    1568.98 7.614e-198
factor(S)     4   914.77      16     654.21 1.049e-196
dist          1   612.32      15      41.89 3.503e-135
factor(dist)  3    29.47      12      12.41  1.780e-06
```

- While distance as a variable is highly significant, change in residual deviance = 612.32 on 1 df, the remainder of distance as a factor is still highly significant (deviance = 29.47 on 3 df). Try a quadratic in `dist`.

```
> mob3=glm(freq~factor(F)+factor(S)+dist+I(dist^2)+factor(dist),family=poisson,data=mobDen)
> anova(mob3, test="Chi")

Analysis of Deviance Table

             Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                          24    2489.01
factor(F)     4   920.03      20    1568.98 7.614e-198
factor(S)     4   914.77      16     654.21 1.049e-196
dist          1   612.32      15      41.89 3.503e-135
I(dist^2)     1    24.11      14      17.78  9.101e-07
factor(dist)  2     5.37      12      12.41       0.07
```

- The model with a quadratic in `dist` provides an adequate fit to the data (residual deviance = 17.78 on 14 df). The remainder of distance as a factor (or, equivalently, the cubic and quartic components of `factor(dist)`) is not significant (just, p-value = 0.07). Hence either of the models with distance as a factor or with a quadratic in distance (as a variable) is reasonable. If we were to use the AIC criterion then the model with distance as a factor would be chosen. For the model with the quadratic in `dist`, AIC = $17.779+2\times(25-14) = 39.779$ (or 180.5699 from R), whereas for the model with distance as a factor, AIC = $12.414+2\times(25-12) = 38.414$ (or 179.2045 from R).

- Should also look at the residuals from various models to check that there is no one cell that needs to be omitted, as was the case with the UK data. This should have been done earlier, but I know that nothing of any real interest comes from doing this.

```
> mob4=glm(freq~factor(F)+factor(S),family=poisson,data=mobDen)
> matrix(resid(mob4), 5, 5)

           [,1]       [,2]       [,3]      [,4]      [,5]
[1,]  7.00220323  3.5289762  0.0789374 -4.329624 -3.897144
[2,]  3.55848799  9.5285854  2.2113592 -5.360075 -7.359953
[3,] -0.08142686  0.6850295  6.2892528 -1.854427 -6.026086
[4,] -4.09082769 -4.3022494 -2.7607313  4.558341  1.104113
[5,] -3.18901581 -8.2956378 -7.1004516  1.252638  9.587359
```

- No one cell stands out as having a residual much larger than all of the others.

```
> mob5=glm(freq~factor(F)+factor(S)+diag,family=poisson,data=mobDen)
> matrix(resid(mob5), 5, 5)

          [,1]      [,2]       [,3]      [,4]      [,5]
[1,]  4.5721291  3.381310  0.4036384 -3.999222 -3.834613
[2,]  3.2077492  3.527857  3.7953673 -3.720404 -6.601678
[3,]  0.1382537  2.192838 -1.2300245  2.725940 -3.266087
[4,] -3.6555374 -2.423791  2.0568082 -2.765010  5.021879
[5,] -3.1026791 -7.405318 -4.1943965  5.092395  2.017614
```

- Again, no one cell stands out as having a residual much larger than all of the others.

```
> mob6=glm(freq~factor(F)+factor(S)+dist,family=poisson,data=mobDen)
> matrix(resid(mob6), 5, 5)

          [,1]       [,2]       [,3]       [,4]       [,5]
[1,]  1.1338724  0.6267074  0.3126952 -2.0820190 -1.2603541
[2,] -0.3548975  0.7036676  0.9331878 -0.8105233 -1.6112992
[3,]  0.1228416 -0.1269085 -1.0751210  1.0231011  0.4717768
[4,] -1.2240886  0.5105106  0.8171778 -1.8787486  1.9461111
[5,]  0.3349659 -3.0205617 -0.3088648  2.5225500 -1.2446066
```

- None of the residuals here is especially large (largest is -3.02).

```
> mob7=glm(freq~factor(F)+factor(S)+dist+I(dist^2),family=poisson,data=mobDen)
> matrix(resid(mob7), 5, 5)

           [,1]       [,2]         [,3]         [,4]       [,5]
[1,]  0.78757599 -0.1184735  0.071202111 -1.288931415  0.1701461
[2,] -1.18098186  1.1286103 -0.007343324 -0.666292954  0.1008993
[3,] -0.03874405 -0.9515770  0.242669718 -0.008563984  0.5470252
[4,] -0.21817923  0.6175566 -0.187013368 -0.449474652  0.5307199
[5,]  2.22542905 -1.7014284 -0.217984584  1.234992803 -0.8329644
```

- One residual out of 25 with magnitude greater than 2, but less than 3, is nothing to get excited about.

```
> mob8=glm(freq~factor(F)+factor(S)+factor(dist),family=poisson,data=mobDen)
> matrix(resid(mob8), 5, 5)

             [,1]       [,2]        [,3]        [,4]       [,5]
[1,]  0.902255167 -0.1136013  0.09789786 -0.78027844 -0.9726934
[2,] -1.188845882  1.0166538 -0.13850620 -0.80017890  0.9461067
[3,] -0.002477728 -1.0375477  0.37186875 -0.05981154  0.4681128
[4,]  0.387812167  0.5256886 -0.23435827 -0.38876054  0.4152440
[5,]  0.776955719 -1.0421046 -0.24947689  1.18356582 -0.7906777
```

- All of the residuals here are quite small.

```
> summary(mob8)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.18885  -0.78028  -0.05981   0.46811   1.18357

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.66990    0.16927  15.773  < 2e-16 ***
factor(F)2     1.21692    0.14685   8.287  < 2e-16 ***
factor(F)3     1.68129    0.14338  11.726  < 2e-16 ***
factor(F)4     1.69478    0.14507  11.683  < 2e-16 ***
factor(F)5     1.51598    0.14786  10.253  < 2e-16 ***
factor(S)2     0.66626    0.13094   5.088 3.61e-07 ***
factor(S)3     1.29329    0.12432  10.403  < 2e-16 ***
factor(S)4     1.50829    0.12505  12.062  < 2e-16 ***
factor(S)5     1.36945    0.12837  10.668  < 2e-16 ***
factor(dist)1 -0.47552    0.04467 -10.645  < 2e-16 ***
factor(dist)2 -1.21517    0.06634 -18.318  < 2e-16 ***
factor(dist)3 -2.42556    0.16324 -14.859  < 2e-16 ***
factor(dist)4 -2.72900    0.36509  -7.475 7.73e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2489.011  on 24  degrees of freedom
Residual deviance:   12.414  on 12  degrees of freedom
AIC: 179.20

Number of Fisher Scoring iterations: 4
```

- The only parameter estimates of interest are those associated with distance (`dist`). The levels of dist go from 0 to 4 and the parameter for the first level (0), which refers to the main diagonal, has been put equal to zero (by `contr.treatment`). The tendency for sons to end up in occupational categories the same as, or more similar to, that of their father is indicated by the increasing, negative estimates for levels of the dist.