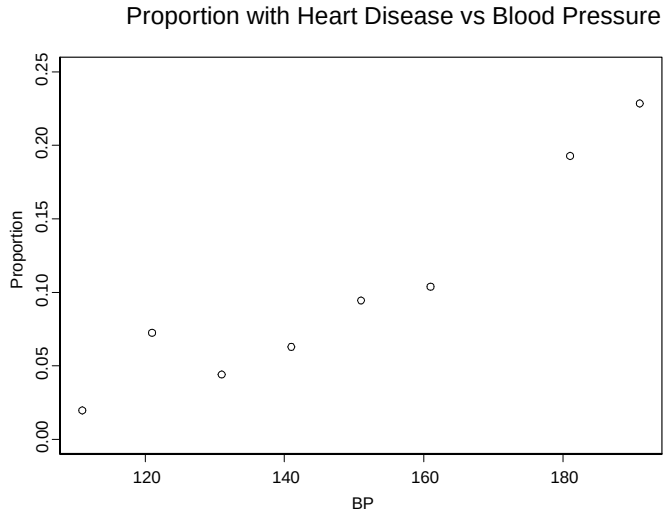
M3S12 BIOSTATISTICS - EXERCISES 4
SOLUTIONS

1. The suitable GLM is the logistic regression: the eight categories are regarded as containing (conditionally) independent Binomially distributed data (giving a product binomial likelihood). Let $\pi_0, ..., \pi_7$ be the eight conditional probabilities

$$\pi_k = P\left(\text{Heart Disease Present}|\text{BP is in category } k\right) = P\left(F|E_k\right) \qquad k = 0, 1, ..., 7.$$

The baseline category can be decided upon arbitrarily; without loss of generality, let $k = 0$ correspond to the < 117 mbHg exposure category. The plot below indicates how the probability depends on blood pressure.



Three models can be fitted:

- M_0 the null model - a one parameter model
- M_1 the model with blood pressure fitted as a continuous predictor - a two parameter model
- M_2 the saturated model - an eight parameter model

The following SPLUS code implements an analysis using the logistic GLM:

```
> bp.cat <- c(1:8)
>bp.bp<-c(111,121,131,141,151,161,181,191)
>bp.y<-c(3,17,12,16,12,8,16,8)
>bp.n<-c(153,235,272,255,127,77,83,35)
>bp.data<-data.frame(bp.cat,bp.bp,bp.y,bp.n)
>names(bp.data)<-c(''BPCAT'',''BP'',''Y'',''N'')
>options(contrasts=c(''contr.treatment'',''contr.poly''))
>fit0.bp<-glm(cbind(Y,N)~1,data=bp.data,family=binomial)
>fit1.bp<-glm(cbind(Y,N)~BP,data=bp.data,family=binomial)
>fit2.bp<-glm(cbind(Y,N)~factor(BPCAT),data=bp.data,family=binomial)
>summary(fit0.bp)
```

```
                Value Std. Error   t value
(Intercept) -2.598656  0.1115164 -23.30289

    Null Deviance: 30.02257 on 7 degrees of freedom
Residual Deviance: 30.02257 on 7 degrees of freedom

> summary(fit1.bp)
Coefficients:
                  Value  Std. Error   t value
(Intercept) -5.95737817 0.693641000 -8.588561
         BP  0.02345621 0.004624166  5.072529

    Null Deviance: 30.02257 on 7 degrees of freedom
Residual Deviance: 5.593988 on 6 degrees of freedom

> summary(fit2.bp)
Coefficients:
                    Value Std. Error   t value
   (Intercept) -3.9318256  0.5829682 -6.744494
factor(BPCAT)2  1.3054535  0.6347683  2.056583
factor(BPCAT)3  0.8109302  0.6533466  1.241195
factor(BPCAT)4  1.1631508  0.6373959  1.824848
factor(BPCAT)5  1.5725452  0.6565511  2.395160
factor(BPCAT)6  1.6674618  0.6912589  2.412210
factor(BPCAT)7  2.2855737  0.6437392  3.550465
factor(BPCAT)8  2.4559191  0.7024410  3.496264

    Null Deviance: 30.02257 on 7 degrees of freedom
Residual Deviance: 0 on 0 degrees of freedom
```

From this analysis it is evident that there is a significant influence on the rate of heart disease of blood pressure. Model M_1 fits significantly better than the null model M_0 (change in deviance of $30.02257 - 5.593988 \approx 24.41$ on a change in degrees of freedom of 1), and model M_1 appears to fit adequately (dispersion estimate $5.593988/6 \approx 1$) In model M_1, the linear predictor takes the form

$$\beta_0 + \beta_1 \times BP$$

and $\widehat{\beta}_1 = 0.0235$ ($e.s.e = 0.005$), so the rate of heart disease increases with increasing BP, and the adequacy of fit indicates that the increase is linear on the logit scale.

2. (a) In the pooled table

	E	E'	TOTAL
F	45	20	65
F'	40	40	80
TOTAL	85	60	145

so

$$\log \widehat{\psi} = \log\left(\frac{n_{11}n_{22}}{n_{12}n_{21}}\right) = \log\left(\frac{45\times 40}{40\times 20}\right) = 0.811$$

$$e.s.e.\left(\log \widehat{\psi}\right) = \sqrt{\frac{1}{n_{11}}+\frac{1}{n_{21}}+\frac{1}{n_{12}}+\frac{1}{n_{22}}} = 0.350$$

and a 95% interval for $\log \widehat{\psi}$ is

$$0.811 \pm 1.96 \times 0.350 = (0.126 : 1.500)$$

and a 95% interval for $\widehat{\psi}$ is

$$\left(e^{0.126} : e^{1.500}\right) = (1.134 : 4.464)$$

(b) For the individual tables, using the same formulae

$$\widehat{\psi}_S = 3.5 \qquad \widehat{\psi}_{NS} = 1.0$$

$$\text{SMOKERS} \quad : \quad 95\% \text{ CI for } \psi_S \text{ is } (1.047 : 11.689)$$

$$\text{NON-SMOKERS} \quad : \quad 95\% \text{ CI for } \psi_{NS} \text{ is } (0.375 : 2.664)$$

and so for smokers, there is some evidence of increased risk of bladder cancer, whereas for non-smokers there is no evidence of increased risk. Now

$$\widehat{\psi}_{NS} < \widehat{\psi} < \widehat{\psi}_S$$

so there is no evidence of **confounding** in the strict sense, where the direction of effect is altered/reversed by aggregating the two individual tables, as in Simpson's paradox - we could only infer confounding in this sense if

$$\widehat{\psi} < \widehat{\psi}_{NS}, \widehat{\psi}_S \qquad \text{or} \qquad \widehat{\psi} > \widehat{\psi}_{NS}, \widehat{\psi}_S$$

However, there is evidence of **interaction** or **effect modification**, as $\widehat{\psi}_{NS} \neq \widehat{\psi}_S$.

(c) It is better to report smoking-specific odds-ratios because of the modification; this does not average/aggregate over tables.

(d) Relative risks (risk ratios): Strictly, the relative risks and risk differences, cannot be computed for this case-control study. In this context, the rare disease hypothesis may be appropriate, so the usual approximation

$$\frac{\pi_1}{\pi_0} \approx \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \psi$$

may be appealed to, and the calculations for each of the ψs above may be used.

Ignoring the case-control aspect, and assuming that the study is actually a cohort study, we have the usual estimates for the measures of effect:

- RELATIVE RISK: π_1/π_0

$$\text{CRUDE (POOLED TABLE)} \quad : \quad \widehat{RR} = \frac{\widehat{\pi}_1}{\widehat{\pi}_0} = \frac{45/85}{20/60} = 1.59$$

$$\text{SMOKERS} \quad : \quad \widehat{RR}_S = \frac{\widehat{\pi}_{S1}}{\widehat{\pi}_{S0}} = \frac{35/55}{5/15} = 1.91$$

$$\text{NON-SMOKERS} \quad : \quad \widehat{RD}_{NS} = \frac{\widehat{\pi}_{NS1}}{\widehat{\pi}_{NS0}} = \frac{10/30}{15/45} = 1.00$$

- RISK DIFFERENCE: $\pi_1 - \pi_0$

$$\text{CRUDE (POOLED TABLE)} \quad : \quad \widehat{RD} = \widehat{\pi}_1 - \widehat{\pi}_0 = \frac{45}{85} - \frac{20}{60} = 0.20$$

$$\text{SMOKERS} \quad : \quad \widehat{RD}_S = \widehat{\pi}_{S1} - \widehat{\pi}_{S0} = \frac{35}{55} - \frac{5}{15} = 0.30$$

$$\text{NON-SMOKERS} \quad : \quad \widehat{RD}_{NS} = \widehat{\pi}_{NS1} - \widehat{\pi}_{NS0} = \frac{10}{30} - \frac{15}{45} = 0.00$$

Again, there is no suspicion of confounding as

$$\widehat{RR}_{NS} \quad < \quad \widehat{RR} < \widehat{RR}_S$$

$$\widehat{RD}_{NS} \quad < \quad \widehat{RD} < \widehat{RD}_S$$

(that is, the conclusion from the pooled analysis is not inconsistent with the smoking-specific analysis), but there is effect modification, as exposure E is only harmful (increases the mortality rate) in the smoking sub-group.

(e) For the case-control study, need to fit logistic regression in the rows of the 2×2 table (although an estimate of ψ is also available from a logistic regression in the columns). In the usual notation

$$\gamma_0 = P\left(E|F'\right) \qquad \gamma_1 = P\left(E|F\right)$$

and the logistic regression model has

$$\text{CONTROLS} \quad : \quad \log\left(\frac{\gamma_0}{1-\gamma_0}\right) = \mu$$

$$\text{CASES} \quad : \quad \log\left(\frac{\gamma_1}{1-\gamma_1}\right) = \mu + \alpha$$

so that $\psi = \exp\{\alpha\}$. In the smoking subtable, $\widehat{\psi}_S = 3.5$, so that $\widehat{\alpha}_S = \log 3.5 = 1.253$, and the parameter μ_S is the log-odds on exposure for the controls, so that

$$\widehat{\mu}_S = \log\left(\frac{n_{21}}{n_{22}}\right) = \log 2 = 0.693$$

(f) In the equivalent cohort study, we have the 40 cases, and an entire (smoking) population of size 300,000. For this study, denote the cell entries in the 2×2 table $\left\{n_{ij}^C, i, j = 1, 2\right\}$, where $n_{11}^C = 35, n_{12}^C =$

5. In the usual notation, let S denote inclusion in the study, so that

$$\text{CONTROLS} \quad : \quad P\left(S|E \cap F'\right) = P\left(S|E' \cap F'\right) = \phi$$

$$\text{CASES} \quad : \quad P\left(S|E \cap F\right) = P\left(S|E' \cap F\right) = 1$$

and we estimate ϕ by

$$\widehat{\phi} = \frac{30}{300000} = 0.0001.$$

Now,

$$\frac{P\left(E \cap F \cap S\right)}{P\left(E \cap F' \cap S\right)} = \frac{P(E)P(F|E)P(S|E \cap F)}{P(E)P(F'|E)P(S|E \cap F')} = \frac{\pi_1}{1-\pi_1}\frac{1}{\phi} \tag{1}$$

$$\frac{P\left(E' \cap F \cap S\right)}{P\left(E' \cap F' \cap S\right)} = \frac{P(E')P(F|E')P(S|E' \cap F)}{P(E')P(F'|E')P(S|E' \cap F')} = \frac{\pi_0}{1-\pi_0}\frac{1}{\phi} \tag{2}$$

and hence the logistic regression in the cohort study (that is, in the columns of the table) is

$$\text{UNEXPOSED} \quad : \quad \log\left(\frac{\pi_0}{1-\pi_0}\right) = \beta$$

$$\text{CASES} \quad : \quad \log\left(\frac{\pi_1}{1-\pi_1}\right) = \beta + \alpha$$

Now

$$\frac{P\left(E \cap F \cap S\right)}{P\left(E \cap F' \cap S\right)} \text{ is estimated by } \frac{n_{11}}{n_{21}}$$

$$\frac{P\left(E' \cap F \cap S\right)}{P\left(E' \cap F' \cap S\right)} \text{ is estimated by } \frac{n_{12}}{n_{22}}$$

and hence

$$\log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\phi \text{ is estimated by } \log\left(\frac{n_{11}}{n_{21}}\right) \quad \text{(from (1)}$$

$$\log\left(\frac{\pi_0}{1-\pi_0}\right) - \log\phi \text{ is estimated by } \log\left(\frac{n_{12}}{n_{22}}\right) \quad \text{(from (2)}$$

so that

$$\log\left(\frac{\pi_1}{1-\pi_1}\right) \text{ is estimated by } \log\left(\frac{n_{11}}{n_{21}}\right) + \log\widehat{\phi}$$

$$\log\left(\frac{\pi_0}{1-\pi_0}\right) \text{ is estimated by } \log\left(\frac{n_{12}}{n_{22}}\right) + \log\widehat{\phi}.$$

Here

$$\log\widehat{\phi} = \log\left(\frac{30}{300000}\right) = -9.210$$

so that

$$\log\left(\frac{\pi_1}{1-\pi_1}\right) \text{ is estimated by } \log\left(\frac{35}{20}\right) + (-0.9210) = -8.651$$

$$\log\left(\frac{\pi_0}{1-\pi_0}\right) \text{ is estimated by } \log\left(\frac{5}{10}\right) + (-0.9210) = -9.903$$

Hence

$$\widehat{\beta} = -9.903 \qquad \widehat{\alpha} = 1.253.$$

Risk:

$$P\left(F|E'\right) = \pi_0 = \frac{\exp\{\beta\}}{1+\exp\{\beta\}} = 5.00 \times 10^{-5}$$

Relative Risk:

$$\frac{P\left(F|E\right)}{P\left(F|E'\right)} = \frac{\pi_1}{\pi_0} = \frac{\dfrac{\exp\{\beta+\alpha\}}{1+\exp\{\beta+\alpha\}}}{\dfrac{\exp\{\beta\}}{1+\exp\{\beta\}}} = 3.496$$

Note that, here, the relative risk and the odds ratio are very nearly equal; thus the rare disease assumption, and consequent approximations may be valid here (incidence of disease in smoking population is $40/300040 \approx 1 \times 10^{-4}$)

3. See SPLUS analysis below

```
>case<-c(0,0,0,1,3,1,0,3,10,20,1,1,1,0,0,0,5,11,21,42,3,8,3,5,3,1,7,19,34,31)
>control<-c(25,13,8,4,2,106,175,153,165,155,25,10,11,4,1,79,142,119,130,96,12,
10,7,1,2,39,73,58,67,50)
>midata<-cbind(case,control)
>options(contrasts=c(''contr.treatment'',''contr.poly''))
>a<-c(1:5)
>age<-factor(rep(a,6))
>s<-c(1:3)
>i<-rep(10,3)
>smoke<-factor(rep(s,i))
>o<-c(1,1,1,1,1,2,2,2,2,2)
>oc<-factor(rep(o,3))
#Models
#Null
>mod0<-glm(midata~1,family=''binomial'')
#Age
>mod1<-glm(midata~age,family=''binomial'')
#Smoke
>mod2<-glm(midata~smoke,family=''binomial'')
#Age+Smoke
>mod3<-glm(midata~age+smoke,family=''binomial'')
#Age*SMoke=Age+Smoke+Age.Smoke
>mod4<-glm(midata~age*smoke,family=''binomial'')
```

```
#################################################################

> summary(mod0)
Call: glm(formula = midata ~1, family = ''binomial'')
Coefficients:
                   Value   Std.Error   t value
(Intercept)      -2.007468 0.1130119 -17.76333

    Null Deviance: 307.5811 on 29 degrees of freedom
Residual Deviance: 307.5811 on 29 degrees of freedom

#################################################################

> summary(mod1)
Call: glm(formula = midata ~age, family = ''binomial'')
Coefficients:
               Value   Std.Error   t value
(Intercept) -3.8641650 0.4101667 -9.420962
        age2 0.8614021 0.4664718  1.846633
        age3 1.6001525 0.4450282  3.595620
        age4 2.2106428 0.4301349  5.139417
        age5 2.7356997 0.4261522  6.419537

    Null Deviance: 307.5811 on 29 degrees of freedom
Residual Deviance: 182.9462 on 25 degrees of freedom

#################################################################

> summary(mod2)
Call: glm(formula = midata ~smoke, family = ''binomial'')
Coefficients:
               Value  Std.Error t value
(Intercept) -3.054493 0.1657507 -18.428232
      smoke2 1.036343 0.2031837   5.100522
      smoke3 2.025500 0.1984439  10.206914

    Null Deviance: 307.5811 on 29 degrees of freedom
Residual Deviance: 184.5141 on 27 degrees of freedom

#################################################################
```

```
> summary(mod3)
Call: glm(formula = midata ~age + smoke, family = ''binomial'')
Coefficients:
                Value   Std.Error t value
(Intercept) -5.0055843 0.4456276 -11.232663
       age2  0.8074202 0.4735044   1.705201
       age3  1.5765143 0.4519820   3.488002
       age4  2.1968863 0.4371293   5.025712
       age5  2.8659886 0.4342755   6.599471
     smoke2  1.1063510 0.2084054   5.308648
     smoke3  2.1657990 0.2072610  10.449620

    Null Deviance: 307.5811 on 29 degrees of freedom
Residual Deviance: 53.25249 on 23 degrees of freedom

##############################################################

> summary(mod4)
Call: glm(formula = midata ~age * smoke, family = ''binomial'')
Coefficients:
               Value   Std. Error  t value
(Intercept) -4.8751973  1.003810 -4.8566956
       age2 -7.7127728 23.966252 -0.3218181
       age3  0.8924052  1.160680  0.7688641
       age4  2.1431939  1.050933  2.0393253
       age5  2.9544457  1.028339  2.8730268
     smoke2  0.2308064  1.420299  0.1625055
     smoke3  2.3296661  1.130151  2.0613758
 age2smoke2  9.1250427 23.990917  0.3803541
 age3smoke2  1.3693579  1.564551  0.8752403
 age4smoke2  0.6478797  1.472806  0.4398948
 age5smoke2  0.8529038  1.449557  0.5883893
 age2smoke3  8.5475137 23.973517  0.3565398
 age3smoke3  0.5697812  1.295232  0.4399068
 age4smoke3 -0.1536087  1.189292 -0.1291598
 age5smoke3 -0.8337977  1.172916 -0.7108759

    Null Deviance: 307.5811 on 29 degrees of freedom
Residual Deviance: 36.67926 on 15 degrees of freedom

##############################################################
```

```
Deviance Summary:
-----------------
             Deviance   DF Dev. Diff. DF Diff
NULL         307.5811   29 - -
AGE          182.9462   25 124.63        4
SMOKE        184.5141   27 123.07        2
AGE+SMOKE     53.2525   23 254.33        6
AGE*SMOKE     36.6793   15 270.90       14


################################################################

Comparing AGE+SMOKE with AGE*SMOKE:

Dev. Diff. = 53.25249-36.67926
DF Diff    = 14-6 = 8
Chisquared(8) distn, 0.95 point: 15.50731
=> SIGNIFICANTLY BETTER FIT.

################################################################

#Stepwise, Deviance-based Model Selection
> step.glm(mod4)
Start: AIC= 66.6793
 midata ~age * smoke
Single term deletions
Model:
midata ~age * smoke
scale: 1
 Df Sum of Sq RSS Cp
 <none> 43.95637 73.95637
age:smoke 8 9.79258 53.74895 67.74895
Step: AIC= 67.2525
 midata ~age + smoke
Call:
glm(formula = midata ~age * smoke, family = ''binomial'')
Degrees of Freedom: 30 Total; 15 Residual
Residual Deviance: 36.67926
################################################################
```

Thus best model is

$$AGE * SMOKE$$

as no smaller model fits as adequately as this one. For this model, the dispersion estimate is

$$\frac{36.6793}{15} > 1$$

so there is some evidence of overdispersion.

4. This question considers how to summarize the results of a series of studies via their odds ratios, and how to test to see whether there is a common odds ratio across the whole series, Let E_i, F_i denote the exposure and incidence events for study i, and

$$\pi_{1i} = P\left(F_i|E_i\right) \qquad \pi_{0i} = P\left(F_i|E_i'\right)$$

and

$$\xi_i = \log \psi_i = \log\left(\frac{\pi_{1i}/(1-\pi_{1i})}{\pi_{0i}/(1-\pi_{0i})}\right).$$

For example, in study 1,

	E	E'	TOTAL
F	59	88	147
F'	3844	3834	7678
TOTAL	3903	3922	7825

and

$$\widehat{\xi}_i = \log \widehat{\psi}_i = \log\left(\frac{59 \times 3834}{88 \times 3844}\right) = -0.402$$

$$e.s.e\left(\widehat{\xi}_i\right) = \sqrt{\frac{1}{59} + \frac{1}{88} + \frac{1}{3844} + \frac{1}{3834}} = 0.1698$$

and so on. See the SPLUS analysis below:

```
>#Q4.StudyData

>x<-c(59,13,60,5,1,25,43,1,2,10,18,32,20)
>m<-c(3903,1721,8700,186,193,1048,233,68,45,49,534,416,419)
>y<-c(88,22,109,20,6,36,52,3,1,21,34,48,39)
>n<-c(3922,1706,8654,194,196,1004,219,63,42,48,529,424,465)

>#EffectsEstimates
>pi1<-x/m
>pi0<-y/n

#RELATIVERISK
>rr<-pi1/pi0

#RISKDIFFERENCE
>delta<-pi1-pi0

#LOG-ODDSRATIO
>xi<-log((pi1/(1-pi1))/(pi0/(1-pi0)))
```

```
#SUMMARY
> round(cbind(pi1, pi0, rr, delta, xi), 4)
        pi1    pi0    rr    delta     xi
 [1,] 0.0151 0.0224 0.6737 -0.0073 -0.4024
 [2,] 0.0076 0.0129 0.5858 -0.0053 -0.5402
 [3,] 0.0069 0.0126 0.5475 -0.0057 -0.6081
 [4,] 0.0269 0.1031 0.2608 -0.0762 -1.4257
 [5,] 0.0052 0.0306 0.1693 -0.0254 -1.8022
 [6,] 0.0239 0.0359 0.6653 -0.0120 -0.4199
 [7,] 0.1845 0.2374 0.7772 -0.0529 -0.3191
 [8,] 0.0147 0.0476 0.3088 -0.0329 -1.2090
 [9,] 0.0444 0.0238 1.8667  0.0206  0.6455
[10,] 0.2041 0.4375 0.4665 -0.2334 -1.1097
[11,] 0.0337 0.0643 0.5245 -0.0306 -0.6775
[12,] 0.0769 0.1132 0.6795 -0.0363 -0.4265
[13,] 0.0477 0.0839 0.5691 -0.0361 -0.6024


#####################################################################
> #Aggregated Log-Odds Ratio Quantities: Test of Homogeneity
> w.xi <- (1/x + 1/(m - x) + 1/y + 1/(m - y))^(-1)
> xi.hat <- sum(w.xi * xi)/sum(w.xi)
> Q.xi <- sum(w.xi * (xi - xi.hat)^2)
> Q.xi
[1] 9.581888
> qchisq(0.95, 12)
[1] 21.02601
#=> NO EVIDENCE FOR A DIFFERENCE IN LOG-ODDS RATIOS ACROSS STUDIES

#Z-test for z.xi
z.xi <- (xi.hat) * sqrt(sum(w.xi))
> z.xi
[1] -6.899617
> #=>EVIDENCE FOR A TREATMENT EFFECT: xi.hat > 0, so there is a BENEFICIAL TREATMENT EFFECT

#####################################################################
#Aggregated Risk Differences Quantities
#Care with Weights (reciprocal of sum of variances)
> w.delta <- ((pi1 * (1 - pi1))/m + (pi0 * (1 - pi0))/n)^(-1)
> delta.hat <- sum(w.delta * delta)/sum(w.delta)
> Q.delta <- sum(w.delta * (delta - delta.hat)^2)
> Q.delta
[1] 28.23193
> qchisq(0.95, 12)
[1] 21.02601
#=> EVIDENCE FOR A DIFFERENCE IN RISK DIFFERENCES ACROSS STUDIES
```