MATH 523 - EXERCISES 3- SOLUTIONS

1 A grouped-data logistic regression model seems most appropriate: to fit BP first as a factor predictor, then as a continuous covariate, the following code may be used:

```
BP<-c(117,126,136,146,156,166,186,200)
Present<-c(3,17,12,16,12,8,16,8)
Absent<-c(153,235,272,255,127,77,83,35)

B<-gl(8,1)

BP.mod1<-glm(cbind(Present,Absent)~B,family="binomial")
BP.mod2<-glm(cbind(Present,Absent)~BP,family="binomial")
```

Results are as follows: for the factor predictor

```
> summary(BP.mod1)

Call: glm(formula = cbind(Present, Absent) ~ B, family = "binomial")

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.9318     0.5830  -6.744 1.54e-11 ***
B2            1.3055     0.6348   2.057 0.039731 *
B3            0.8109     0.6534   1.241 0.214543
B4            1.1632     0.6374   1.825 0.068030 .
B5            1.5725     0.6566   2.395 0.016615 *
B6            1.6675     0.6913   2.412 0.015858 *
B7            2.2856     0.6438   3.550 0.000385 ***
B8            2.4559     0.7025   3.496 0.000472 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

    Null deviance:  3.0023e+01  on 7  degrees of freedom
Residual deviance: -1.2657e-14  on 0  degrees of freedom
AIC: 48.701
```

and for the continuous covariate (after making the pragmatic decision to take the upper ends of the BP categories as the value of the continuous covariate)

```
> summary(BP.mod2)

Call: glm(formula = cbind(Present, Absent) ~ BP, family = "binomial")

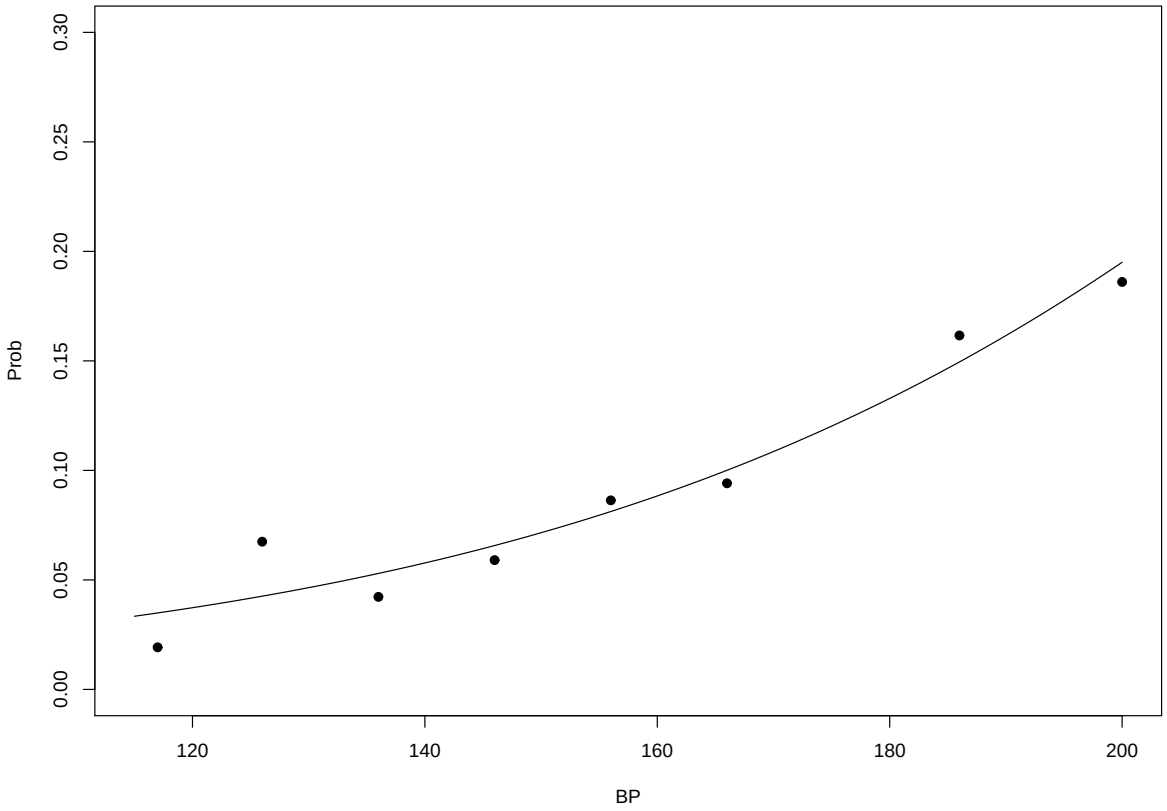
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1640  -0.5452  -0.1667   0.2470   1.8089

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.999723   0.700251  -8.568  < 2e-16 ***
BP           0.022910   0.004507   5.083 3.72e-07 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.0226  on 7  degrees of freedom
Residual deviance:  5.7386  on 6  degrees of freedom
AIC: 42.44
```

There is clear evidence that the incidence of heart disease **increases** with increasing blood pressure; this follows from inspection of the coefficients and their associated standard errors and p-values. The following plot indicates that the continuous covariate model seems to be quite adequate.



2 (a) Taking the log-odds with the treatment group in the denominator, the estimates and standard errors are

Study	$\widehat{\xi}_k$	s.e.$(\widehat{\xi}_k)$	t
1	-0.4024	0.1698	-2.3698
2	-0.5402	0.3515	-1.5369
3	-0.6081	0.1615	-3.7657
4	-1.4257	0.5111	-2.7893
5	-1.8022	1.0850	-1.6611
6	-0.4199	0.2642	-1.5895
7	-0.3191	0.2318	-1.3764
8	-1.2090	1.1683	-1.0348
9	0.6455	1.2440	0.5189
10	-1.1097	0.4586	-2.4198
11	-0.6775	0.2982	-2.2721
12	-0.4265	0.2395	-1.7811
13	-0.6024	0.2837	-2.1231

Thus there seems to be some evidence that there is an effect of treatment. Given the definition of ψ_k, if $\xi_k < 0$, then the treatment is protective, as

$$\frac{\pi_{11k}/\pi_{01k}}{\pi_{10k}/\pi_{00k}} < 1 \qquad \Longleftrightarrow \qquad \pi_{11k} < \pi_{10k}$$

(b) Using the code provided, $Q_\xi = 9.5812$, whereas $\chi^2_{12}(0.95) = 21.026$, so there is no evidence to suggest that the odds ratios are different across studies.

(c) The Wald-type statistic in this case is

$$\frac{\overline{\xi}}{\sqrt{W_\xi^{-1}}} = -6.8996$$

(d) The standard error of δ_k is

$$\sqrt{\frac{\widehat{\pi}_{11k}(1-\widehat{\pi}_{11k})}{m_{1k}} + \frac{\widehat{\pi}_{10k}(1-\widehat{\pi}_{10k})}{m_{0k}}}$$

and thus we have Working through the calculations, we end up with $Q_\delta = 28.2319$, which

Study	$\widehat{\delta}_k$	s.e.$(\widehat{\delta}_k)$	t
1	-0.0073	0.0031	-2.3869
2	-0.0053	0.0034	-1.5539
3	-0.0057	0.0015	-3.8210
4	-0.0762	0.0248	-3.0675
5	-0.0254	0.0133	-1.9055
6	-0.0120	0.0075	-1.5945
7	-0.0529	0.0384	-1.3783
8	-0.0329	0.0305	-1.0776
9	0.0206	0.0387	0.5333
10	-0.2334	0.0919	-2.5405
11	-0.0306	0.0132	-2.3125
12	-0.0363	0.0202	-1.7975
13	-0.0361	0.0165	-2.1843

after comparing with $\chi^2_{12}(0.95) = 21.026$ is clearly seen to be a significant result. Hence the hypothesis of a constant risk difference is **rejected.**

3 Here we essentially have a $2 \times 2 \times 3$ study (LBW by Smoking Status by Race), so here we have essentially three 2×2 tables.

(a) For the Mantel-Haenzsel analysis, we can test for a common odds ratio relating LBW to Smoking across the three race groups. Coding White = 1, Black = 2, Other = 3, we have for the log odds ratios Here, the Q_ξ statistic is 3.8537, which is to be compared against

Study	$\widehat{\xi}_k$	s.e.$(\widehat{\xi}_k)$	t
1	1.7505	0.5983	2.9259
2	1.1939	0.8412	1.4194
3	0.2231	0.6492	0.3437

$\chi^2_{12}(0.95) = 5.9912$, so there is no evidence of heterogeneity. The subsequent Wald test for the hypothesis that the common log odds ratio is equal to zero is based on the statistic

$$\frac{\overline{\xi}}{\sqrt{W_\xi^{-1}}} = 2.596$$

indicating a significantly positive log odds ratio. This indicates that the smoking status influences low birth weight in an approximately equal fashion across all race groups, and that smokers are significantly more likely to give birth to low birth weight babies.

(b) For the logistic regression modelling, an analysis would be based on the saturated model

$$\text{SMOKING} * \text{RACE}$$

and submodels. We can compute the following deviance figures:

Model	p	Dev.	rdf
Null	1	17.854	5
SMOKE	2	16.226	4
RACE	3	8.292	3
SMOKE+RACE	4	8.256	2
SMOKE*RACE	6	0.000	0

Dropping the two-way interaction leads to a deviance change of 8.256 on two degrees of freedom, which is significant as $\chi^2_{12}(0.95) = 5.9912$; this implies that the additive (independence) model is inadequate. The interaction model has

$$\log\left(\frac{\pi_{1jk}}{1-\pi_{1jk}}\right) = \beta_0 + \beta_j^{(S)} + \beta_k^{(R)} + \gamma_{jk}^{(SR)}$$

with the usual identifiability constraints. We have

$$\log\left(\frac{\pi_{12k}/(1-\pi_{12k})}{\pi_{11k}/(1-\pi_{11k})}\right) = \beta_2^{(S)} + \gamma_{2k}^{(SR)}$$

so $\gamma_{2k}^{(SR)} = 0$ means a common odds ratio across strata of k (RACE).

Comparing the SMOKE+RACE model with the simpler SMOKE model (omitting the race effect, and thus assuming an odds ratio of 1 in all race groups), it is clear that this simplification is also rejected.

4 To investigate the association between oral contraceptive use and myocardial infarction, we can fit the models involving the three factors `oc` (oral contraceptive use, two levels: Yes, No), `age` (Age, five levels), and `smoke` (smoking status, three levels).

Beginning with the saturated model, `oc*smoke*age` we attempt to omit the three way interaction first and proceed from there;

Model	p	$\triangle$ Dev.	$\triangle$ rdf	p-value	Conclusion
`oc*age*smoke`	30	-	-		
`-oc.age.smoke`	22	9.200	8	0.3257	Not Significant
`-oc.age`	18	5.297	4	0.2582	Not significant
`-oc.smoke`	16	3.653	2	0.1610	Not significant
`-smoke.age`	15	15.824	1	0.0451	Significant

On this basis, we cannot drop the final two-way interaction, so by backwards selection analysis of deviance, we conclude that

```
oc + age + smoke + smoke.age = oc + smoke*age
```

This model concludes that the effect of OC use is independent of stratification by smoking and age. However, this method of dropping single terms is not always stable. Consider the following sequence of fits, where we decide to drop `smoke.age` before `oc.smoke`; this might be considered, as neither remaining two-way interaction is significant.

Model	p	$\triangle$ Dev.	$\triangle$ rdf	p-value	Conclusion
`oc*age*smoke`	30	-	-		
`-oc.age.smoke`	22	9.200	8	0.3257	Not Significant
`-oc.age`	18	5.297	4	0.2582	Not significant
`-smoke.age`	10	13.727	8	0.0891	Not Significant
`-oc.smoke`	8	5.740	2	0.0567	Not significant

which allows us to **omit** the final two way interaction, yielding the model

`oc + age + smoke`

A direct comparison of the two models needs to be carried out; the change of deviance is 15.814 on 8 degrees of freedom, which yields a p-value of 0.04512. This we can retain the model including the final two-way interaction.

```
1  > summary(mod.d3)
2
3  Call: glm(formula = midata ~ oc + age + smoke + age:smoke, family = "binomial")
4
5  Coefficients:
6                Estimate Std. Error z value Pr(>|z|)
7  (Intercept)   -4.03847    1.01675  -3.972 7.13e-05 ***
8  oc2           -1.19641    0.26706  -4.480 7.47e-06 ***
9  age2         -17.02077 2794.92497  -0.006  0.99514
10 age3           1.14894    1.16511   0.986  0.32407
11 age4           2.44769    1.05704   2.316  0.02058 *
12 age5           3.26505    1.03488   3.155  0.00161 **
13 smoke2         0.14224    1.42287   0.100  0.92037
14 smoke3         2.22984    1.13478   1.965  0.04941 *
15 age2:smoke2   18.74152 2794.92519   0.007  0.99465
16 age3:smoke2    1.40498    1.56765   0.896  0.37013
17 age4:smoke2    0.74659    1.47573   0.506  0.61292
18 age5:smoke2    0.98075    1.45251   0.675  0.49954
19 age2:smoke3   18.01739 2794.92504   0.006  0.99486
20 age3:smoke3    0.60522    1.30055   0.465  0.64168
21 age4:smoke3   -0.06961    1.19431  -0.058  0.95352
22 age5:smoke3   -0.75593    1.17797  -0.642  0.52106
23 ---
24 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
25
26 (Dispersion parameter for binomial family taken to be 1)
27
28     Null deviance: 307.581  on 29  degrees of freedom
29 Residual deviance:  18.150  on 14  degrees of freedom
30 AIC: 123.86
31
32 Number of Fisher Scoring iterations: 17
```

- Note the large standard errors for the `age2` categories; this arises from the fact that in the 30-34 age group, non smokers, there are no incidences of MI in either OC group. This means that we have essentially one parameter fewer that can be fitted.
- there is an increasing rate of MI with OC use (coefficient of `oc2` is -1.19641 (0.267), and `oc2` corresponds to no OC use). Also, there is an increasing rate of MI with age (lines 9-12) and smoking status (lines 13-14).
- Overall, the deviance of 18.150 on 14 degrees of freedom indicates a reasonable fit for the logistic regression.