Department of Statistics

Course STATS 330

Model Answer for Assignment 4, 2003

Note: Your answer for this assignment was **NOT** expected to be in the form of a report. It should have had 6 sections corresponding to the 6 questions.

Question 1.

> *Set up a 6-level factor, P, whose levels correspond to the 6 distinguishable combinations of the parents' eye colour. Note that the eye colours of fathers and mothers were not separately recorded. Suggested R-code for doing this is given at the end of the assignment.*

The 6 levels of factor P are:

Both parents have light eyes (P1 in code below)
Both parents have hazel eyes (P2 in code below)
Both parents have dark eyes (P3 in code below)
One has light, the other hazel (P4 in code below)
One has light, the other dark (P5 in code below)
One has hazel, the other dark (P6 in code below)

R code to define a suitable P was given in the assignment (assuming you read in the supplied data set)

```
galton.df<-read.table(file.choose(),header=T)
P<-character(78)
for(i in 1:78){
v<-galton.df[i,1:3]
P[i]<-if(all(v==c(2,0,0))) "P1" else
   if(all(v==c(0,2,0))) "P2" else
   if(all(v==c(0,0,2))) "P3" else
   if(all(v==c(1,1,0))) "P4" else
   if(all(v==c(1,0,1))) "P5" else
   if(all(v==c(0,1,1))) "P6" else NA
}
P<-factor(P)
```

Question 2

Fit a logistic regression to the proportions of children in each family that have light-coloured eyes, using P as the explanatory variable. Are there any outliers? Delete any clearly discrepant points. (You have an outlier budget of 2.)

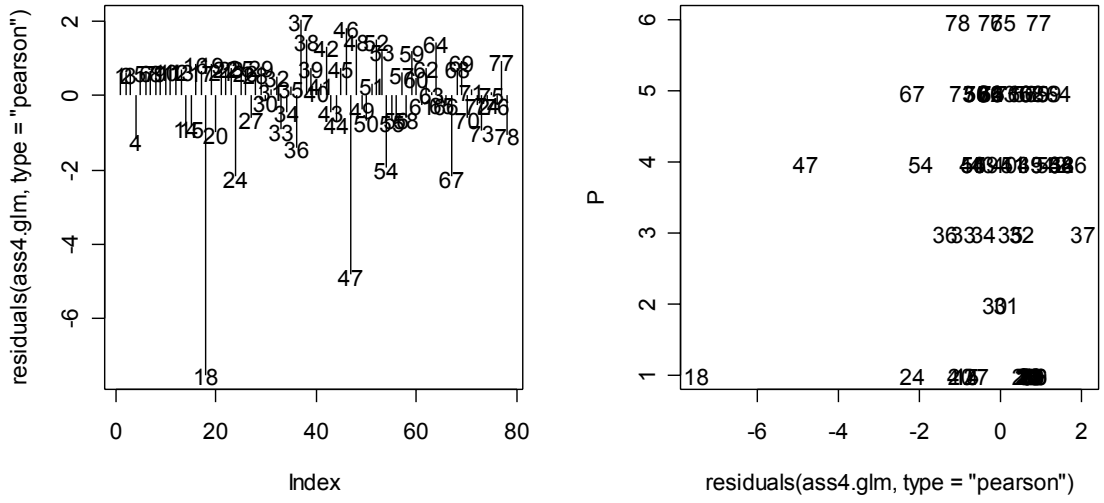
We can fit a logistic regression (a logistic "one-way anova") using the factor P. We first make a data frame, using P and G (see question 5 for G).

```
ass4.df<-data.frame(P=P, G=G, galton.df[,7:8])
ass4.glm<-glm(cbind(r,n-r)~P,family=binomial, data=ass4.df)
```

We plot deviance residuals using the code

```
par(mfrow=c(1,2))
plot(residuals(ass4.glm,type="deviance"),type="h")
text(residuals(ass4.glm,type="deviance"))

plot(residuals(ass4.glm,type="deviance"), P,type="n")
text(residuals(ass4.glm,type="deviance"), P)
```



Clearly points 18 and 47 are atypically large, so we delete these. The group one residuals are also rather strange: they appear to largely lie in two separate clusters. A possible explanation for this is that one of the clusters has been misclassified: perhaps it belongs in another group. Still, since we can't check this, we take it as genuine. In a real application, we would check this out more thoroughly.

Question 3

Refit the model to the remaining data. Compute the fitted probabilities for all 6 eye-colour combinations.

```
model2<-glm(cbind(r,n-r)~P,family=binomial,
        subset=(1:78) [-c(18,47)], data=ass4.df)
fitted.probs<-predict(model2,
        newdata=data.frame(P=levels(P)),type="response")
> fitted.probs
[1] 0.9686096 0.6000000 0.2444444 0.7841727 0.5436242
0.4857143

Note that, since the fitted probabilities are just the
group proportions, we could also have used the code

tapply(r[-c(18,47)],P[-c(18,47)],sum)/
      tapply(n[-c(18,47)],P[-c(18,47)],sum)

       P1        P2        P3        P4        P5        P6
0.9686099 0.6000000 0.2444444 0.7841727 0.5436242 0.4857143
```

Now make a 3×3 table of all the 9 possible eye-colour combinations of the 2 parents, distinguishing mother from father.

Probability of a child with light coloured eyes

Father's eyes	**Mother's eyes**		
	Light	*Hazel*	*Dark*
Light	*0.969*	*0.784*	*0.544*
Hazel	*0.784*	*0.600*	*0.486*
Dark	*0.544*	*0.486*	*0.244*

Question 4

Using these assumptions, transfer your fitted probabilities to the 3×3 table. Display the table and comment on any marked trends and patterns. What you think these data say about the inheritance of eye colour?

The trends in the table are clear: as parental eye colour goes from light to hazel to dark, the chance of the child having light-coloured eyes decreases. The effect of both parents having light-coloured eyes doesn't double the probability that results from only one parent having light eyes: if you have at least one parent with light eyes, your chance of

having light eyes is more than 50%. Thus, a light-coloured gene tends to be stronger than a dark-coloured one.

Question 5.

Now set up the factor G (as you did for P in Question 1) for all the distinguishable eye colour combinations of the grandparents. How many levels does this factor potentially have? How many of these levels are actually in the data set? Include your R code and print out the whole data frame with variables P, G, r and n.

There are 15 possible arrangements of grandparents' eyecolour, assuming we only count how *many* grandparents have certain coloured eyes, and not *which* grandparents. The 15 possibilities can be generated by the R-code

```
for(i in 0:4){
   for(j in 0:(4-i)) print(c(i,j,4-i-j))}
[1] 0 0 4
[1] 0 1 3
[1] 0 2 2
[1] 0 3 1
[1] 0 4 0
[1] 1 0 3
[1] 1 1 2
[1] 1 2 1
[1] 1 3 0
[1] 2 0 2
[1] 2 1 1
[1] 2 2 0
[1] 3 0 1
[1] 3 1 0
[1] 4 0 0
```

We can create the factor using code similar to that used in Question 1:

```
G<-character(78)
for(i in 1:78){
v<-galton.df[i,4:6]
G[i]<-if(all(v==c(0,0,4))) "G1" else
   if(all(v==c(0,1,3))) "G2" else
   if(all(v==c(0,2,2))) "G3" else
   if(all(v==c(0,3,1))) "G4" else
   if(all(v==c(0,4,0))) "G5" else
   if(all(v==c(1,0,3))) "G6" else
   if(all(v==c(1,1,2))) "G7" else
   if(all(v==c(1,2,1))) "G8" else
   if(all(v==c(1,3,0))) "G9" else
   if(all(v==c(2,0,2))) "G10" else
```

```
   if(all(v==c(2,1,1))) "G11" else
   if(all(v==c(2,2,0))) "G12" else
   if(all(v==c(3,0,1))) "G13" else
   if(all(v==c(3,1,0))) "G14" else
   if(all(v==c(4,0,0))) "G15" else NA
}
G=factor(G,levels=paste("G",1:15,sep=""))
# note the levels have to be changed, otherwise 10 comes
# before 2 etc
```

In actual fact, only 12 of the levels are present in the data:

```
> table(G)
G
 G1  G2  G3  G4  G5  G6  G7  G8  G9 G10 G11 G12 G13 G14 G15
  0   1   0   0   1   5   4   1   1  12   7   5  16   8  17
```

G1, G3, and G4 are not in the data. The whole data frame is constructed by

```
ass4.df<-data.frame(P=P, G=G, galton.df[,7:8])
```

```
> ass4.df
    P   G  n  r
1  P1  G1  6  6
2  P1  G1  6  6
3  P1  G1  6  6
4  P1  G1  6  5
5  P1  G1  7  7
6  P1  G1  7  7
7  P1  G1  7  7
8  P1  G1  7  7
9  P1  G1  7  7
10 P1  G1  8  8
11 P1  G1  8  8
12 P1  G1  8  8
13 P1  G1  8  8
14 P1  G1  8  7
15 P1  G1  8  7
16 P1  G1 12 12
17 P1  G2  7  7
18 P1  G2 10  4
19 P1  G2 12 12
20 P1  G3  7  6
21 P1  G3  8  8
22 P1  G3  9  9
23 P1  G3  9  9
24 P1  G3  9  7
```

25	P1	G3	10	10
26	P1	G4	7	7
27	P1	G4	10	9
28	P1	G5	6	6
29	P1	G5	10	10
30	P2	G5	7	4
31	P2	G6	8	5
32	P3	G3	6	2
33	P3	G6	9	1
34	P3	G7	6	1
35	P3	G7	11	3
36	P3	G8	6	0
37	P3	G8	7	4
38	P4	G2	6	6
39	P4	G2	7	6
40	P4	G2	8	6
41	P4	G2	9	7
42	P4	G2	11	10
43	P4	G3	9	6
44	P4	G3	11	7
45	P4	G4	7	6
46	P4	G4	9	9
47	P4	G4	11	1
48	P4	G6	6	6
49	P4	G6	6	4
50	P4	G6	8	5
51	P4	G6	9	7
52	P4	G5	6	6
53	P4	G5	10	9
54	P4	G9	9	4
55	P4	G8	8	5
56	P5	G1	7	3
57	P5	G3	6	4
58	P5	G3	7	3
59	P5	G3	8	6
60	P5	G3	8	5
61	P5	G3	8	4
62	P5	G3	9	6
63	P5	G3	9	5
64	P5	G6	6	5
65	P5	G6	6	3
66	P5	G6	8	4
67	P5	G6	10	2
68	P5	G6	14	9
69	P5	G5	7	5
70	P5	G10	7	3
71	P5	G8	7	4

```
72 P5  G7  8  4
73 P5  G7  8  3
74 P5 G11  6  3
75 P6  G6  6  3
76 P6  G5  9  4
77 P6  G7 13  8
78 P6 G12  7  2
```

Question 6

Add this factor to your model. What evidence is there for a grandparent effect over and above the parental effect?

Fitting a model with grandparents has its problems: we don't have enough data to fit a full factorial model with every P, G combination represented:

```
> table(G,P)
     P
G     P1 P2 P3 P4 P5 P6
  G1   0  0  0  0  0  0
  G2   0  0  0  0  1  0
  G3   0  0  0  0  0  0
  G4   0  0  0  0  0  0
  G5   0  0  0  0  0  1
  G6   0  0  2  0  2  1
  G7   0  0  2  1  1  0
  G8   0  0  0  0  1  0
  G9   0  0  0  1  0  0
  G10  0  1  1  4  5  1
  G11  2  1  0  2  1  1
  G12  2  0  0  3  0  0
  G13  6  0  1  2  7  0
  G14  3  0  0  5  0  0
  G15 16  0  0  0  1  0
```

Most of the cells are empty. However, R will fit a probability to every non-empty cell, and can perform a sensible test of the hypothesis that G adds nothing to the model (assuming that P is already present).

```
model3<-glm(cbind(r,n-r)~P*G,family=binomial,
       subset=(1:78)[-c(18,47)],data=ass4.df)
```

To test the hypothesis of a zero grandparent effect, we compare model 2 ie the model fitted by

```
glm(cbind(r,n-r)~P,family=binomial,
               subset=(1:78) [-c(18,47)], data=ass4.df)
```

using anova:

```
> anova(model2,model3)
Analysis of Deviance Table

Model 1: cbind(r, n - r) ~ P
Model 2: cbind(r, n - r) ~ P * G
  Resid. Df Resid. Dev Df Deviance
1        70     72.224
2        46     46.676 24   25.548
```

We assess the significance (i.e. get a p-value) by typing

```
> 1-pchisq(25.548,24)
[1] 0.3764903
```

Since the p-value 0.3765 is greater than 0.05, there is no evidence of a grandparent effect: the model involving the parental effects alone is OK.