Example H Effect of process and purity index on fault occurrence

Description of data (Fictitious data based on a real investigation)

Minor faults occur irregularly in an industrial process and as an aid to their diagnosis, the following experiment was done. Batches of raw material were selected and each batch was divided into two equal sections: for each batch, one of the sections was processed by the standard method and the other by a slightly modified method, in which the temperature at one stage is reduced. Before processing, a purity index was measured for the whole batch of material. For the product from each section of material it was recorded whether the minor faults did or did not occur. Results for 22 batches are given in Table H.1.

Table H.1 Occurrence of faults in 22 batches

Purity index	Standard process	Modified process	Purity index	Standard process	Modified process
7.2	NF	NF	6.5	NF	F
6.3	F	NF	4.9	F	F
8.5	F	NF	5.3	F	NF
7.1	NF	F	7.1	NF	F
8.2	F	NF	8.4	F	NF
4.6	F	NF	8.5	NF	F
8.5	NF	NF	6.6	F	NF
6.9	F	F	9.1	NF	NF
8.0	NF	NF	7.1	F	NF
8.0	F	NF	7.5	NF	F
9.1	NF	NF	8.3	NF	NF

General considerations

The data here are so limited that in practice very detailed analysis would hardly be justified. The unusual features of the data are the pairing, combined with the availability of a quantitative explanatory variable; the response variable is binary.

Rather than plunge straight into the fitting of a relatively complex models, it is wise to start by simple analysis, first ignoring the explanatory variable and then examining the effect of that variable by simple graphs or tables. Maximum-likelihood fitting of various models can then follow, with a simple basis having been laid for understanding the answers.

The analysis

If we ignore purity index, a standard technique for assessing matched-pair data with binary responses involves the 14 pairs with the mixed response, these being split between 5 'NF,F' and 9 'F,NF'. This suggests a higher chance of fault in the standard process. The null hypothesis of process difference is tested via the binomial distribution with 14 trials and probability 1/2; the two-sided level obtained via a normal approximation with continuity correction is

$$2\Phi\left(-\frac{|5-7|-1/2}{\sqrt{14\times1/2\times1/2}}\right) = 2\Phi(-0.8017837) = 0.423 \qquad \text{(H.1)}$$

so that the apparent process effect is entirely consistent with chance fluctuations.

To examine the effect of purity index on its own, a graph, Fig.H.1, of grouped proportion of faults versus purity index shows that most of the faults occur on batches of low purity index.

To investigate both effects, we fit a linear logistic model by maximum likelihood. To fit from first principles, rather than a package such as GLIM, the model is taken in the approximately orthogonal form that for the ith batch with purity index x_i,

$$\Pr(fault|standard\ process) = \frac{\exp(\alpha + \Delta + \beta(x_i - \bar{x}))}{1 + \exp(\alpha + \Delta + \beta(x_i - \bar{x}))}$$
$$\Pr(fault|modified\ process) = \frac{\exp(\alpha - \Delta + \beta(x_i - \bar{x}))}{1 + \exp(\alpha - \Delta + \beta(x_i - \bar{x}))} \qquad (H.2)$$

where $\bar{x}$ is the mean of the x_i . It is provisionally assumed that responses are independent when the purity index is in the model.

Table H.2 compares the results of fitting the model (H.2) and of reduced models with $\Delta = 0$, with $\beta = 0$ and with $\beta = \Delta = 0$.

Table H.2 Fitting of linear logistic models

Model	No. of parameters		Maximum log likelihood	Estimates ± std.errors	
mean	3	α	-26.40593	$\hat{\alpha}$	-0.4124 ± 0.3334
process difference		Δ		$\hat{\Delta}$	-0.4322 ± 0.3358
purity index		β		$\hat{\beta}$	-0.6042 ± 0.2838
mean	2	α	-29.01005	$\hat{\alpha}$	-0.3811 ± 0.3128
process difference		Δ		$\hat{\Delta}$	-0.3811 ± 0.3128
mean	2	α	-27.26125	$\hat{\alpha}$	-0.3592 ± 0.3257
purity index		β		$\hat{\beta}$	-0.5795 ± 0.2763
mean	1	α	-29.76714	$\hat{\alpha}$	-0.3677 ± 0.3066

The last two models correspond to two and one simple binomial distributions. The results confirm the simpler analyses. The data are not consistent with constant probability (*Cox here refers to the last model*) of fault; χ^2 with 2 degrees of freedom is

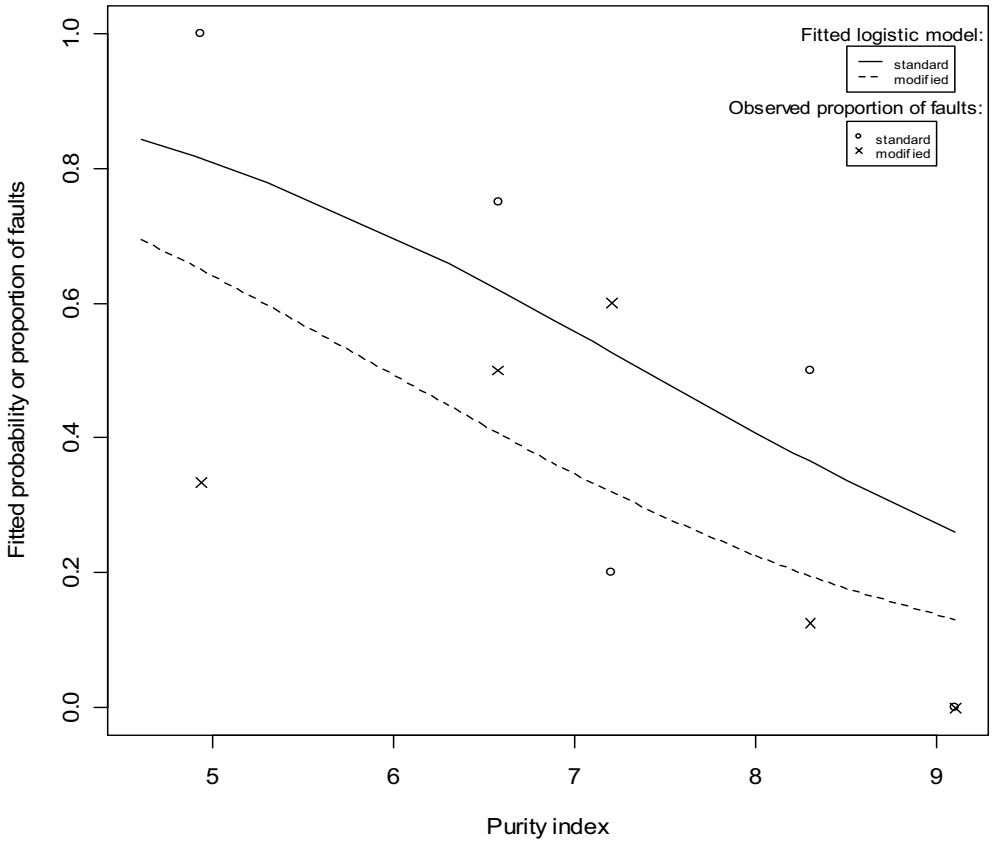
$$2\left(l(\hat{\beta}) - l(\tilde{\beta})\right) = 2(-26.40593 + 29.76714) = 6.722 \sim \chi^2(2)$$

and is just significant at 5% $\left(since\ 6.722 > 5.99 = \chi^2(0.95, 2)\right)$. The estimate of the logistic process difference is 2Δ=2(0.4322)= 0.864 with a standard error of 0.672. Thus a wide range of differences, including $\Delta =$ 0 is consistent with data. Positive means that the standard process has the higher probability of fault. A trend with purity index is moderately well established; $\hat{\beta} =$ -0.6042, with

a standard error of 0.284, the 2-sided p-value (3.3%) is less than 5% and the trend is in the direction (*Cox refers to the negative sign of* $\hat{\beta}$) expected on general grounds.

To interpret the parameters and to check the adequacy of the model, Fig.H.1, shows the fitted models, i.e. the curves (H.2) with α, Δ, β replaced by $\hat{\alpha}, \hat{\Delta}, \hat{\beta}$. The figure also shows the observed proportions of faults based on a grouping into five sets with roughly constant purity index in each set. The plot exposes the paucity of data, revealed also by very wide confidence limits for β and Δ; if desired, these limits, too, could be illustrated graphically.

Fig.H.1 Fitted logistic models



Further points and exercises

(i) Find the exact binomial probability corresponding to Eqn (H.1).
(ii) Compare tests of $\beta = 0$ and of $\Delta = 0$ from maximum likelihood estimates and their standard error, with those based on maximized log likelihood.
(iii) The model (H.2) assumes independence of the two responses in a pair. How can this be tested and, if necessary, dependence allowed for?
(iv) An alternative to the linear logistic model is to record 1 for fault, 0 for no fault, to fit a linear model for expressed response, i.e. for the probability of a fault. Compare this with the results of Table H.2; under what circumstances would this approach be expected to give appreciably different answers form the linear logistic model?

Appendix: R Code

```
######################
# R code for Example H
######################

#-------------------------------------
# fitting 3 logistic regression models
# listed in Table H.2
# Find estimates and standard errors
#-------------------------------------

P1 = c(72,72,63,63,85,85,71,71,82,82,46,46,85,85,69,69,80,80,80,80,91,91)
P2 = c(65,65,49,49,53,53,71,71,84,84,85,85,66,66,91,91,71,71,75,75,83,83)
purity = 0.1 * c(P1, P2)
process = rep(c('1', '2'), 22)      # '1' for standard, '2' for modified
F1 = c(0,0,1,0,1,0,0,1,1,0,1,0,0,0,1,1,0,0,1,0,0,0)
F2 = c(0,1,1,1,1,0,0,1,1,0,0,1,1,0,0,0,1,0,0,1,0,0)
fault = c(F1,F2)

options(contrasts=c("contr.sum","contr.poly"))

# fit the full model

h1.glm = glm(fault ~ I(purity-mean(purity)) + factor(process),
             family = binomial)
summary(h1.glm)
```

```
> summary(h1.glm)

Call:
glm(formula = fault ~ I(purity - mean(purity)) + factor(process),
    family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5381  -0.9187  -0.6321   1.0687   1.8622

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.4124     0.3334  -1.237   0.2161
I(purity - mean(purity))  -0.6042     0.2838  -2.129   0.0333 *
factor(process)1       0.4322     0.3358   1.287   0.1981
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.534  on 43  degrees of freedom
Residual deviance: 52.812  on 41  degrees of freedom
AIC: 58.812

Number of Fisher Scoring iterations: 4
```

```
# fit the reduced model in which 'purity index' is removed

h2.glm = update(h1.glm, . ~ . - I(purity - mean(purity)))
summary(h2.glm)
```

```
> summary(h2.glm)

Call:
glm(formula = fault ~ factor(process), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1774  -0.9508  -0.8752   1.1774   1.5134
```

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.3811     0.3128  -1.218    0.223
factor(process)1   0.3811     0.3128   1.218    0.223

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.534  on 43  degrees of freedom
Residual deviance: 58.020  on 42  degrees of freedom
AIC: 62.02

Number of Fisher Scoring iterations: 4
```

```
# fit the reduced model in which 'purity index' is included

h3.glm = glm(fault ~ I(purity-mean(purity)), family=binomial)
summary(h3.glm)
```

```
> summary(h3.glm)

Call:
glm(formula = fault ~ I(purity - mean(purity)), family = binomial)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.710  -0.899  -0.716   1.179   1.648

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.3952     0.3257  -1.213   0.2250
I(purity - mean(purity))  -0.5795     0.2763  -2.097   0.0360 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.534  on 43  degrees of freedom
Residual deviance: 54.522  on 42  degrees of freedom
AIC: 58.522

Number of Fisher Scoring iterations: 4
```

```
# fit the reduced model when both 'purity index' and 'process' is removed

h4.glm = glm(fault ~ 1, family=binomial)
summary(h4.glm)
```

```
> summary(h4.glm)

Call:
glm(formula = fault ~ 1, family = binomial)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.026  -1.026  -1.026   1.337   1.337

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3677     0.3066  -1.199    0.230

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.534  on 43  degrees of freedom
Residual deviance: 59.534  on 43  degrees of freedom
AIC: 61.534

Number of Fisher Scoring iterations: 4
```

```
#-------------------------------------------------
# calculating maximum log-likelihood in Table H.2
# using model (H.1)
#-------------------------------------------------

P1 = c(72,72,63,63,85,85,71,71,82,82,46,46,85,85,69,69,80,80,80,80,91,91)
P2 = c(65,65,49,49,53,53,71,71,84,84,85,85,66,66,91,91,71,71,75,75,83,83)
```

```
purity = 0.1*c(P1, P2)
identity = rep(c(1,-1),22)
F1 = c(0,0,1,0,1,0,0,1,1,0,1,0,0,0,1,1,0,0,1,0,0,0)
F2 = c(0,1,1,1,1,0,0,1,1,0,0,1,1,0,0,0,1,0,0,1,0,0)
fault = c(F1,F2)
tab = data.frame(purity,process,identity,fault)
```

```
> tab
   purity process identity fault
1     7.2       1        1     0
2     7.2       2       -1     0
3     6.3       1        1     1
4     6.3       2       -1     0
5     8.5       1        1     1
6     8.5       2       -1     0
7     7.1       1        1     0
8     7.1       2       -1     1
9     8.2       1        1     1
10    8.2       2       -1     0
11    4.6       1        1     1
12    4.6       2       -1     0
13    8.5       1        1     0
14    8.5       2       -1     0
15    6.9       1        1     1
16    6.9       2       -1     1
17    8.0       1        1     0
18    8.0       2       -1     0
19    8.0       1        1     1
20    8.0       2       -1     0
21    9.1       1        1     0
22    9.1       2       -1     0
23    6.5       1        1     0
24    6.5       2       -1     1
25    4.9       1        1     1
26    4.9       2       -1     1
27    5.3       1        1     1
28    5.3       2       -1     0
29    7.1       1        1     0
30    7.1       2       -1     1
31    8.4       1        1     1
32    8.4       2       -1     0
33    8.5       1        1     0
34    8.5       2       -1     1
35    6.6       1        1     1
36    6.6       2       -1     0
37    9.1       1        1     0
38    9.1       2       -1     0
39    7.1       1        1     1
40    7.1       2       -1     0
41    7.5       1        1     0
42    7.5       2       -1     1
43    8.3       1        1     0
44    8.3       2       -1     0
```

```
# maximum log-likelihood for the full model

alpha1 = coefficients(h1.glm)[1]
beta1 = coefficients(h1.glm)[2]
delta1 = coefficients(h1.glm)[3]
num1 = exp(alpha1 + delta1*identity + beta1*(purity - mean(purity)))
prob1 = num1 / (1 + num1)
loglik1 = sum(fault * log(prob1/(1 - prob1)) + log(1 - prob1))
loglik1                          # -26.40593

# maximum log-likelihood for reduced model in which 'purity' is removed

alpha2 = coefficients(h2.glm)[1]
delta2 = coefficients(h2.glm)[2]
num2 = exp(alpha2 + delta2 * identity)
prob2 = num2 / (1 + num2)
loglik2 = sum(fault * log(prob2/(1 - prob2)) + log(1 - prob2))
loglik2                          # -29.01005
```

```
# maximum log-likelihood for reduced model in which 'process' is removed

alpha3 = coefficients(h3.glm)[1]
beta3 = coefficients(h3.glm)[2]
num3 = exp(alpha3 + beta3 *(purity-mean(purity)))
prob3 = num3 / (1 + num3)
loglik3 = sum(fault * log(prob3/(1 - prob3)) + log(1 - prob3))
loglik3                          # -27.26125

# maximum log-likelihood for reduced model in which
# both 'process' and 'purity' deleted

alpha4 = coefficients(h4.glm)[1]
num4 = exp(alpha4)
prob4 = num4 /(1 + num4)
loglik4 = sum(fault * log(prob4/(1-prob4)) + log(1-prob4))
loglik4                          # -29.76714

#---------------------------------
# reproducing Fig. H.1
#---------------------------------

# plot fitted probability vs purity

i = order(purity[process=="2"], h1.glm$fitted.values[process=="2"])
x.mod = purity[process=="2"][i]
y.mod = h1.glm$fitted.values[process=="2"][i]
j = order(purity[process=="1"], h1.glm$fitted.values[process=="1"])
x.std = purity[process=="1"][j]
y.std = h1.glm$fitted.values[process=="1"][j]
plot(purity, fault, type='n', xlab='Purity index',
     ylab='Fitted probability or proportion of faults',
     main='Fig.H.1  Fitted logistic models')
lines(x.mod, y.mod, lty = 2)
lines(x.std, y.std, lty=1)

# plot observed proportion of faults vs purity

k = order(purity[process=="2"], fault[process=="2"], fault[process=="2"])
purity.ord = purity[process=="1"][k]
fault.std.ord = fault[process=="1"][k]
fault.mod.ord = fault[process=="2"][k]

##> purity.ord
## [1] 4.6 4.9 5.3 6.3 6.5 6.6 6.9 7.1 7.1 7.1 7.2 7.5 8.0 8.0 8.2 8.3 8.4 8.5 8.5 8.5 9.1 9.1
##> fault.std.ord
## [1] 1 1 1   1 0 1 1   0 0 1 0 0   0 1 1 0 1 0 0 1   0 0
##> fault.mod.ord
## [1] 0 1 0   0 1 0 1   1 1 0 0 1   0 0 0 0 0 0 1 0   0 0

## Grouping the data into 5 sets with roughly constant purity in each set
## set1: (4.6,4.9,5.3) with average purity 4.933
##       observed proportion of faults for 'standard'=3/3 (since 3 Fs)
##       observed proportion of faults for 'modified'=1/3 (since 2 NFs, 1 F)
## set2: (6.3,6.5,6.6,6.9) with average purity 6.575
##       observed proportion of faults for 'standard'=3/4 (since 3 Fs, 1 NF)
##       observed proportion of faults for 'modified'=2/4 (since 2 NFs, 2 Fs)
```

```
## set3: (7.1,7.1,7.1,7.2,7.5) with average purity 7.2
##       observed proportion of faults for 'standard'=1/5 (since 1 F, 4 NFs)
##       observed proportion of faults for 'modified'=3/5 (since 2 NFs, 3 Fs)
## set4: (8.0,8.0,8.2,8.3,8.4,8.5,8.5,8.5) with average purity 8.3
##       observed proportion of faults for 'standard'=4/8 (since 4 F, 4 NFs)
##       observed proportion of faults for 'modified'=1/8 (since 7 NFs, 1 Fs)
## set4: (9.1,9.1) with average purity 9.1
##       observed proportion of faults for 'standard'=0/2 (since 2 NFs)
##       observed proportion of faults for 'modified'=0/2 (since 2 NFs)
## So for 'standard' process, 5 points
##       (4.933,3/3),(6.575,3/4),(7.2,1/5),(8.3,4/8),(9.1,0)
## So for 'modified' process, 5 points
##       (4.933,1/3),(6.575,2/4),(7.2,3/5),(8.3,1/8),(9.1,0)

## plot the 5 points for two processes separately

x.set = c(4.933,6.575,7.2,8.3,9.1)
pr.std = c(3/3,3/4,1/5,4/8,0/2)
pr.mod = c(1/3,2/4,3/5,1/8,0/2)
points(x.set,pr.std,pch=1)
points(x.set,pr.mod,pch=4)

## Add legends

tex = c('standard', 'modified')
text(8.75,1, paste("Fitted logistic model:"), cex=0.8)
legend(8.5, 0.98, tex, lty=c(1,2), cex=0.6)
text(8.55, 0.89, paste("Observed proportion of faults:"), cex=0.8)
legend(8.5, 0.87, tex, pch=c(1,4), cex=0.6)
```

Fig.H.1 Fitted logistic models