

Collected Notes on Solutions to S-Plus worksheets: DRAFT

P.M.E.Altham, Statistical Laboratory, University of Cambridge.

December 12, 2006

Not to be quoted without acknowledgement, please.

P.M.E.Altham@statslab.cam.ac.uk

I originally constructed these draft solutions for our MPhil students in about 1998. I tended to make very limited use of graphics at that time, so now I have more leisure I will add the odd extra graph here and there, and maybe expand the explanation in places.

Notes on solution to worksheet 1.

First, Figure 1 shows the plot of y against x . You will see that although y is certainly an increasing function of x , this dependence is just possibly quadratic rather than linear. Here is the linear regression (the output is slightly edited to save paper).

Call: `lm(formula = y ~ x)`

Coefficients:

	Value	Std. Error	t value	Pr(> t)
(Intercept)	-9.4543	0.6553	-14.4283	0.0000
x	1.6579	0.0753	22.0139	0.0000

Residual standard error: 0.2012 on 19 degrees of freedom

Multiple R-Squared: 0.9623

F-statistic: 484.6 on 1 and 19 degrees of freedom, the p-value is 5.551e-15

Correlation of Coefficients:

(Intercept)
x -0.9978

Our model is

$$y_i = \alpha + \beta x_i + \epsilon_i$$

for $i = 1, \dots, n$ where we assume that $(\epsilon_i, i = 1, \dots, n)$ form a random sample from $N(0, \sigma^2)$.

Thus our fitted line is

$$y = -9.4543(0.6553) + 1.6579(0.0753)x$$

with our estimate of σ^2 as $(0.2012)^2$.

The "t-value" is (estimate/its se), which is referred to the t_{n-2} distribution to check for the significance of that coefficient.

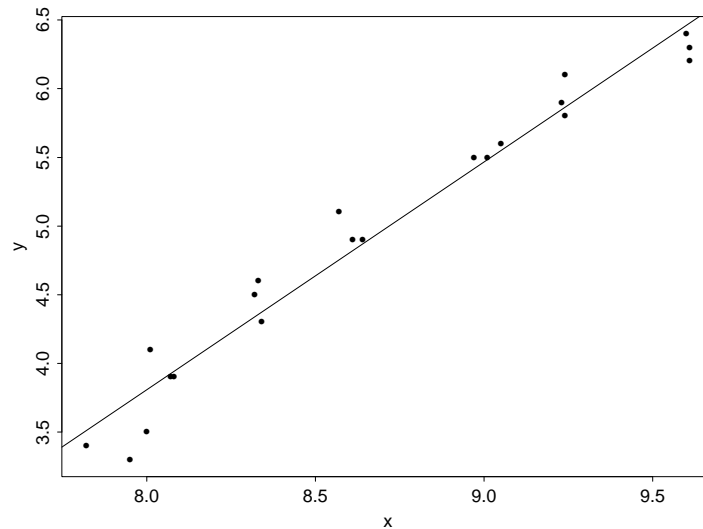


Figure 1: The very first graph

The fit of the straight line is very good: the “Multiple R-Squared” is defined as the “sum of squares due to the regression” divided by “the total sum of squares for y ”.

This is a quantity necessarily between 0 and 1, and so the value of 0.9623 found above shows a very good fit.

The diagnostic plots obtained by

```
plot(teeny,ask=T)
```

provide us with ‘eye-ball’ checks of our basic assumptions that the errors are normal and with constant variance. Since (ϵ_i) are necessarily unobservable, we perform these checks with the natural estimates of (ϵ_i) , namely the *residuals*, which are defined as the (observed values - fitted values) . These checks will be described in more detail later.

The anova table below shows the sum of squares due to the regression, with the residual sum of squares below; these two add to give the “total sum of squares”, which is $\Sigma(y_i - \bar{y})^2$.

Analysis of Variance Table

Response: y

Terms added sequentially (first to last)

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
x	1	19.62028	19.62028	484.6133	5.551115e-15
Residuals	19	0.76924	0.04049		

The “F Value” is the test statistic for testing

$$H_0 : \beta = 0.$$

This “F Value” is necessarily the square of the t-value for β given above.

Now we may wish to check whether a quadratic gives a significantly better fit

than a straight line.

Thus our next task is to fit

$$y_i = \alpha + \beta x_i + \gamma x_i^2$$

for $i = 1, \dots, n$. This is achieved by

```
xx = x*x
```

```
Call: lm(formula = y ~ x + xx)
```

Coefficients:

	Value	Std. Error	t value	Pr(> t)
(Intercept)	-41.5662	9.9392	-4.1821	0.0006
x	9.0430	2.2833	3.9605	0.0009
xx	-0.4227	0.1306	-3.2356	0.0046

Residual standard error: 0.1644 on 18 degrees of freedom

Multiple R-Squared: 0.9761

F-statistic: 368.3 on 2 and 18 degrees of freedom, the p-value is 2.554e-15

Correlation of Coefficients:

	(Intercept)	x
x	-0.9996	
xx	0.9985	-0.9996

Note that

i) $\hat{\gamma} = -0.4227(0.1306)$ showing that we do need a quadratic rather than a line, although the improvement to R^2 is quite marginal,

ii) the coefficient $\hat{\beta}$ has changed dramatically from the previous regression. This is linked with the fact that the parameters β, γ are severely non-orthogonal, due to the fact that for this range of (x_i) , x^2 is almost a linear function of x .

This model is a special case of

$$y = X\beta + \epsilon$$

and here the $n \times 3$ matrix X has two of its columns almost linearly dependent, thus $X^T X$ is nearly a *singular* 3×3 matrix. One consequence of this near-collinearity is that the correlations between the parameter estimates are very close to +1 or to -1.

We can reformulate the quadratic regression to give mutually orthogonal parameters thus:

$$y_i = \alpha_0 P_0(x_i) + \alpha_1 P_1(x_i) + \alpha_2 P_2(x_i) + \epsilon_i$$

for $i = 1, \dots, n$, where

$P_\nu(\cdot)$ is a polynomial of degree ν such that

$$\sum_i P_j(x_i) P_k(x_i)$$

is zero for j, k distinct, and 1 for $j = k$.

Check that with this new formulation

$$y = Z\alpha + \epsilon$$

say, where the matrix Z , which is a linear function of X , satisfies

$$Z^T Z = I$$

the identity matrix.

In this case we are using **orthogonal polynomials**, supplied to us by S-Plus as follows.

```
Call: lm(formula = y ~ poly(x, 2))
```

Coefficients:

	Value	Std. Error	t value	Pr(> t)
(Intercept)	4.9381	0.0359	137.6645	0.0000
poly(x, 2)1	4.4295	0.1644	26.9467	0.0000
poly(x, 2)2	-0.5319	0.1644	-3.2356	0.0046

Residual standard error: 0.1644 on 18 degrees of freedom

Multiple R-Squared: 0.9761

F-statistic: 368.3 on 2 and 18 degrees of freedom, the p-value is 2.554e-15

Correlation of Coefficients:

	(Intercept)	poly(x, 2)1
poly(x, 2)1	0	
poly(x, 2)2	0	0

Note that

- i) reparametrising in this way has no effect on the fit: R^2 remains the same
- ii) the t-value for the quadratic coefficient is exactly the same (-3.2356) whichever parametrisation we use,
- iii) the correlation matrix for $\hat{\alpha}$, which is derived from the inverse of $Z^T Z$, is the identity matrix.

Notes on solution to worksheet 2.

First, a plot of y against x , and then the linear regression of $y = \text{brain-weight}$, on $x = \text{body-weight}$.

You see that y increases with x (no surprise here!)

```
plot(x,y) # to see the graph on the screen in front of you
postscript(file="plot2.ps")
plot(x,y) # to send the graph to the postscript file
dev.off() # to turn off the "current device"
species.lm _ lm(y~x) ; summary(species.lm)
```

```
Call: lm(formula = y ~ x)
```

Coefficients:

	Value	Std. Error	t value	Pr(> t)
(Intercept)	91.0044	43.5526	2.0895	0.0409
x	0.9665	0.0477	20.2778	0.0000

Residual standard error: 334.7 on 60 degrees of freedom

Multiple R-Squared: 0.8727

F-statistic: 411.2 on 1 and 60 degrees of freedom, the p-value is 0

Correlation of Coefficients:

	(Intercept)
x	-0.2176

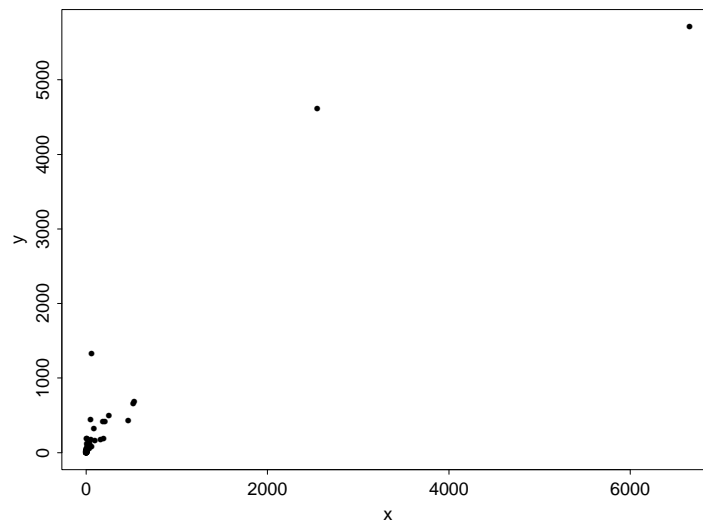


Figure 2: Brain weight versus body weight

The diagnostic plots show that the African elephant is an extreme in the sense that it has a very high Cook's distance relative to the other mammals, so that removing this one mammal could change our fitted regression by a great deal. You could try this by

```
summary(lm(y[-33] ~ x[-33]))
```

However the qq-plot was also unsatisfactory. We now try fitting a power-law, ie

$$y = constant \times x^\beta$$

so that

$$\log(y) = \alpha + \beta \log(x).$$

```
Call: lm(formula = log(y) ~ log(x))
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-1.715	-0.4923	-0.06162	0.436	1.948

```
Coefficients:
```

	Value	Std. Error	t value	Pr(> t)
(Intercept)	2.1348	0.0960	22.2273	0.0000
log(x)	0.7517	0.0285	26.4087	0.0000

```
Residual standard error: 0.6943 on 60 degrees of freedom
```

```
Multiple R-Squared: 0.9208
```

```
F-statistic: 697.4 on 1 and 60 degrees of freedom, the p-value is 0
```

```
Correlation of Coefficients:
```

	(Intercept)
log(x)	-0.3964

Observe that

i) R^2 is better

ii) the diagnostic plots are better.

Notes on combining the Splus output and graphs into your Latex file.

While you are in Splus, you may want to put various results into a file. You can do this by opening an emacs window, and cutting and pasting as you go along, or you may find it easier to ‘sink’ the output to a separate file, thus

```
teeny _ lm(y~x)
summary(teeny)      # displayed on your screen
sink("myresults")
summary(teeny)      # sent to the file named
sink()              # return to the screen
lteeny _ lm(log(y) ~ log(x))
summary(lteeny)
sink("myresults", append=T)
lteeny
sink()
.....and so on
q()
```

At the end of your analysis, you may want to edit the results, ie the S-plus output, and then write a report which incorporates the graph, which you have already put into the file called

plot2.ps

This shows you how to proceed.

```
cp myresults report2.tex
```

Now edit the file report2.tex as follows.

```
\documentclass[12pt,a4]{article}
\usepackage[dvips]{graphicx}
\begin{document}
  Here is our solution to Worksheet 2.
  .....insert numerical results, and commentary, here.

  First we show the graph of  $y$  against  $x$ .
  \begin{figure}
  \centering
  \includegraphics[angle=0, width=0.8\textwidth]{plot2.ps}
  \caption{Brain weight versus body weight}\label{fig:nonlogbb}
  \end{figure}
  You will see that  $y$  is an increasing function of  $x$ .

  ...
\end{document}
```

Note that you can rotate the graph through 90 degrees, if required, by

```
\includegraphics[angle=90, width=0.8\textwidth]{plot2.ps}
```

Suppose your analysis is written up as above in a file called “report2.tex”

Then (in unix) what you must do is

```
latex report2
```

```
xdvi report2
```

This will give you (eventually, after some hard work on the editing) the correct version of your text.

Now
dvips report2
will produce the postscript file
report2.ps
which will refer to the plot2.ps file containing your graph.
ghostview report2.ps
enables you to see the text and graph combined, on your screen.
lp report2.ps
enables you to print it out.
And, better still, to save paper

```
psnup -2 report2.ps | lp
```

Note: if you have several graphs plotted simultaneously in your graphics window, the S-Plus commands given above for the plotting may not be so convenient. So here is an alternative method:

```
par(mfrow=c(2,1))  
plot(x,y) # for first graph  
plot(x,z) # for second graph  
# Now click on 'Print graph' in the graphics window  
q()
```

You will get a hard copy of the (pair of) graphs, and you will be told that the graph has been put into a file ps.out...
This file you can then cp into "plot2.ps", and incorporate into your report as above.

Notes on solution to Worksheet 3.

```
regr _ lm(y~x)
```

has the effect of fitting

$$H_0 : y_{ij} = \alpha + \beta x_i + \epsilon_{ij}$$

for $i = 1, \dots, 5, j = 1, \dots, 3$. You will find that this linear regression fits only moderately well, with a negative slope.

```
potash _ factor(x)
aov(y~ potash)
```

has the effect of a factor declaration, and then fitting the model

$$H_1 : y_{ij} = \alpha + \theta_i + \epsilon_{ij}$$

for $i = 1, \dots, 5, j = 1, \dots, 3$ and then giving the corresponding analysis of variance. We fit H_1 because we want then to test the hypothesis $H_2 : \theta_i = 0, i = 1, \dots, 5$ which is the null hypothesis of no difference between the 5 groups.

Observe

- i) H_0 is a special case of H_1 , so that H_1 is bound to give a better fit than H_0 ,
- ii) the parameters (θ_i) in H_1 are not **identifiable**: we could replace

$$\alpha, (\theta_i)$$

by

$$\alpha - 13.2, (\theta_i + 13.2)$$

for example, and we still have just the same model for the observations. For this reason we need to impose a constraint on the parameters. The standard “glm” constraint is

$$\theta_1 = 0.$$

Unfortunately this is not the default constraint imposed by S-Plus, which uses instead the “Helmert” parametrisation. This makes it very hard to understand what the estimates of (θ_i) are actually telling us.

Worksheet 4 shows you how to change the constraint to the standard glm one. The aov gives us the F-statistic: this is say $((R_2 - R_1)/4)/(R_1/df)$ where R_1, R_2 are the residual ss fitting H_1, H_2 respectively. This is what we use to test H_2 (see the corresponding p-value).

The hypothesis test should only be a relatively small aspect of our modelling, our ‘client’ may well have other questions about the data.

Notes on Solution to Worksheet 4.

General remarks about orthogonal sets of parameters.

Reminder: in fitting

$$E(Y) = X\beta = X_1\beta_1 + X_2\beta_2$$

we say that β_1, β_2 are orthogonal if and only if

$$X_1^T X_2 = 0.$$

Here the matrix X has been partitioned into $(X_1 X_2)$ and β^T into $(\beta_1^T \beta_2^T)$.

Then, if $R_1 =$ resid ss fitting $E(Y) = X_1\beta_1$, and $R_2 =$ resid ss fitting $E(Y) = X_2\beta_2$, and $R =$ resid ss fitting $E(Y) = X\beta$, and $R_0 =$ resid ss fitting $E(Y) = 0$, then

$$R_0 + R = R_1 + R_2.$$

It is sometimes troublesome (for example for the current problem) to check the condition

$$X_1^T X_2 = 0$$

directly, and therefore we may prefer to use the following Lemma.

Lemma

With the model

$$Y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon$$

where $\epsilon \sim N(0, \sigma^2 I)$, the parameters β_1, β_2 are orthogonal if and only if

$$\hat{\beta}_1 = \beta_1^*$$

where $\hat{\beta}_1$ is the lse of β_1 in fitting

$$E(Y) = X\beta = X_1\beta_1 + X_2\beta_2$$

and β_1^* is the lse of β_1 in fitting $E(Y) = X_1\beta_1$.

All of the above can be extended to the model

$$E(Y) = \mu 1 + X\beta$$

where 1 is used as the vector with every element 1.

The present example:

The model we will fit is

$$\omega : p_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

for $i = 1, \dots, 12$ (the country) and $j = 1, \dots, 4$ (the occupation) with the usual assumption that the errors form a random sample from $N(0, \sigma^2)$.

A special feature of this model is that because we have a **balanced design**, that is, each combination (i, j) is observed exactly the same number of times, here once, the group of parameters α is orthogonal to the group of parameters β . We point out various consequences of this orthogonality in the course of presenting the analysis.

This model is of course a special case of

$$Y = X\beta + \epsilon$$

and in this instance the observation vector Y is a vector of 48 elements. The purpose of our preliminary S-Plus commands is to set up the corresponding vectors for COUNTRY and OCC.

The S-Plus function

```
expand.grid()
```

enables us to set up these two vectors to follow the correct (nested) pattern:
thus

```
x
```

```

      Var1 Var2
1 bus/train Den
2  surgeon Den
3 barrister Den
4      MP   Den
5 bus/train Neth
6  surgeon Neth
.....
44      MP   It
45 bus/train Irl
46  surgeon Irl
47 barrister Irl
48      MP   Irl

```

If we impose the standard glm constraints, as suggested we find that when we fit ω as a linear model so that $\alpha_1 = 0 = \beta_1$ then

```
Call: lm(formula = p ~ OCC + COUNTRY)
```

```
.....
Coefficients:
```

	Value	Std. Error	t value	Pr(> t)
(Intercept)	80.8750	2.2951	35.2380	0.0000
OCCsurgeon	4.5833	1.6761	2.7345	0.0100
OCCbarrister	4.0833	1.6761	2.4362	0.0204
OCCMP	6.8333	1.6761	4.0769	0.0003
COUNTRYNeth	-6.7500	2.9031	-2.3251	0.0264
COUNTRYFra	-13.5000	2.9031	-4.6502	0.0001
COUNTRYUK	-16.7500	2.9031	-5.7697	0.0000
COUNTRYBel	-18.7500	2.9031	-6.4586	0.0000
COUNTRYSpa	-20.5000	2.9031	-7.0614	0.0000
COUNTRYPort	-23.0000	2.9031	-7.9225	0.0000
COUNTRYW.Ger	-26.0000	2.9031	-8.9559	0.0000
COUNTRYLux	-28.0000	2.9031	-9.6448	0.0000
COUNTRYGre	-28.0000	2.9031	-9.6448	0.0000
COUNTRYIt	-28.7500	2.9031	-9.9032	0.0000
COUNTRYIrl	-33.5000	2.9031	-11.5394	0.0000

Residual standard error: 4.106 on 33 degrees of freedom

Multiple R-Squared: 0.8908

F-statistic: 19.22 on 14 and 33 degrees of freedom, the p-value is 6.162e-12

Observe that we are comparing all the occupations with the *first* such, namely 'bus/train', and we are comparing all the countries with the *first* such, namely Denmark.

In fact a modest amount of algebra here shows that you can find the least-squares estimators $\hat{\alpha}_i, \hat{\beta}_j$ directly, ie without writing out the design matrix X . Here it is, as a little exercise for you.

Minimise

$$\Sigma \Sigma (p_{ij} - m - a_i - b_j)^2$$

with respect to the parameters $m, (a_i), (b_j)$ subject to the constraints $\Sigma a_i = 0, \Sigma b_j = 0$. Now define

$$m + a_i + b_j = \mu + \alpha_i + \beta_j$$

for all i, j with $\alpha_1 = 0 = \beta_1$, and hence show that $\hat{\alpha}_i =$ (mean for i th row - mean for 1st row), etc.

It is easy to see that if we now minimise $\Sigma \Sigma (p_{ij} - \mu - \alpha_i)^2$, subject to $\alpha_1 = 0$, our estimate of α_i is STILL (mean for i th row - mean for 1st row): thus verifying that the parameters (α_i) and (β_j) are mutually orthogonal.

We would probably not generally want the whole correlation matrix of the parameter estimates printed in full, but we present it here to show one of the consequences of **the orthogonality of the sets of parameters (α_i) to (β_j)** . Thus you will see from this matrix that

$$\text{corr}(\hat{\alpha}_i, \hat{\beta}_j) = 0.$$

Now we could do the

```
aov()
```

command in two possible orders: we demonstrate this below.

```
summary(aov(p~ COUNTRY + OCC))
```

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
COUNTRY	11	4243.917	385.8106	22.88854	0.000000000
OCC	3	291.750	97.2500	5.76944	0.002751829
Residuals	33	556.250	16.8561		

explanation:

The three terms in the above sum add to give the

‘total ss’ (also called total deviance) = resid ss fitting $H_0 : E(p_{ij}) = \mu$.

‘COUNTRY’ = reduction in deviance when we now fit $H_1 : E(p_{ij}) = \mu + \alpha_i$

‘OCC’ = the further reduction in deviance when we now fit $\omega : E(p_{ij}) = \mu + \alpha_i + \beta_j$.

Hence the **sequence** in which the terms above appear in the formula `aov()` is generally important, but here we get the same results whichever way round we put ‘OCC’, ‘COUNTRY’, because they are orthogonal.

```
summary(aov(p~ OCC+ COUNTRY ))
```

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
OCC	3	291.750	97.2500	5.76944	0.002751829
COUNTRY	11	4243.917	385.8106	22.88854	0.000000000
Residuals	33	556.250	16.8561		

```
tapply(p,COUNTRY,mean)
```

Den	Neth	Fra	UK	Bel	Spa	Port	W.Ger	Lux	Gre	It	Irl
84.75	78	71.25	68	66	64.25	61.75	58.75	56.75	56.75	56	51.25

```
tapply(p,OCC,mean)
```

bus/train	surgeon	barrister	MP
60.58333	65.16667	64.66667	67.41667

```
lex2$coefficients
```

```
(Intercept) COUNTRYNeth COUNTRYFra COUNTRYUK COUNTRYBel COUNTRYSpa
```

80.875	-6.75	-13.5	-16.75	-18.75	-20.5
COUNTRYPort	COUNTRYW.Ger	COUNTRYLux	COUNTRYGre	COUNTRYIt	COUNTRYIr1
-23	-26	-28	-28	-28.75	-33.5
OCCsurgeon	OCCbarrister	OCCMP			
4.583333	4.083333	6.833333			

Thus, for example,

$$\hat{\alpha}_2 = -6.75 = 78 - 84.75$$

the mean for Neth minus the mean for Denmark. Orthogonality implies that our estimates for, say, (α_i) are the same whether or not (β_j) are included in the model.

A note on interactions between factors

Suppose that we have the model

$$\Omega : y_{ijk} = \mu_{ij} + \epsilon_{ijk} \text{ for } 1 \leq k \leq n_{ij}$$

and $i = 1, \dots, I$ corresponds to the level of factor A , say, and $j = 1, \dots, J$ corresponds to the level of factor B say. Then Ω is equivalent to

$$\Omega : y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

and it is easily checked that minimising $\sum \sum \sum (y_{ijk} - \mu_{ij})^2$ with respect to (μ_{ij}) gives

$$\hat{\mu}_{ij} = \sum_k y_{ijk} / n_{ij}.$$

We may want to consider the following submodel of Ω , namely

$$\omega : y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}.$$

The model ω corresponds to there being *no interaction* between A, B ; thus we have implicitly defined the parameters (γ_{ij}) in the model Ω above as the *interaction parameters*.

We can test ω against Ω by the usual F-test, based on

$$\frac{(R_\omega - R_\Omega)}{R_\Omega}$$

Note that $\dim(\Omega) - \dim(\omega) = (I - 1)(J - 1)$.

If we then conclude that ω holds, we have a simpler model than Ω , and it will be easier to interpret.

But, if we reject ω in favour of Ω , then we must interpret the resulting interaction between A, B . There are two helpful methods

i) use

`model.tables()`

for summaries of corresponding means, with se's of differences

ii) for each j plot \bar{y}_{ij} against i : this will give you J different 'tracks', which FAIL to be parallel because of the interaction. Discuss this, in a sentence, for your client, with the help of

`interaction.plot()`

Note, if we have factors A, B, C , then

`lm(y~ (A+B+C)^2)`

is interpreted as a command to fit the model

$$\Omega : y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \theta_{ij} + \phi_{jk} + \psi_{ik} + \epsilon_{ijk}$$

ie a model with all main effects (α_i) ... etc and all pairwise interactions (θ_{ij}) ... etc. With our previous choice for ‘options’, we will have the usual ‘corner-point’ glm constraints, ie

$$\alpha_1 = 0, \beta_1 = 0, \gamma_1 = 0,$$

and

$$\theta_{1j} = 0 \text{ for all } j, \theta_{i1} = 0 \text{ for all } i$$

and 2 further sets of constraints.

In general we seek to simplify the model by dropping high-order interactions as far ‘down’ as possible: you may like to try

`step.lm()`

for this. Note that it will not make sense to drop, say, the main effect of A BEFORE dropping a related pairwise interaction, say AB . Well-designed software does not ‘allow’ you to commit such a crime.

With data (y_{ijk}), we could have started the proceedings by

`lm(y~(A+B+C)^3)`

which would be interpreted as fitting

$$\Omega_1 : y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \theta_{ij} + \phi_{jk} + \psi_{ik} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijk}$$

ie a model with all main effects, all pairwise interactions (θ_{ij}) ... etc, AND the 3-factor interactions ($(\alpha\beta\gamma)_{ijk}$). If we only have ONE observation y for each ijk combination, then the model Ω_1 will be a *saturated* model: it contains exactly as many independent parameters as there are observations. Thus we will get a perfect fit, giving residual ss = 0, df = 0, and hence we have no way of estimating σ^2 .

Notes on Solution to Worksheet 5.

The model we will fit is say

$$\omega : y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

for $i = 1, \dots, 10$ and $j = 1, \dots, 5$, except that not all combinations (i, j) are observed.

Here i corresponds to the *row* in the table, ie the author, and j to the *column*, ie the country.

As usual we assume that (ϵ_{ij}) form a random sample from $N(0, \sigma^2)$.

We assume the Glim constraints, $\alpha_1 = 0, \beta_1 = 0$.

We consider various sub-models of ω :

$$H_1 : E(y_{ij}) = \mu + \alpha_i \text{ and}$$

$$H_2 : E(y_{ij}) = \mu + \beta_j \text{ and}$$

$$H_0 : E(y_{ij}) = \mu.$$

Because of the lack of *balance* in this experimental design, we find

i) in fitting ω , $cov(\hat{\alpha}_i, \hat{\beta}_j)$ are non-zero, for some i, j ,

ii) the lse's for, say, (α_i) are not the same for models ω and H_1 ,

iii) the reduction in deviance in moving from H_1 to ω is not the same as the reduction in deviance in moving from H_0 to H_2 .

iv) although we can easily work out the algebraic formulae for, say, the lse's for (α_i) in fitting H_1 , the same is not true when fitting ω , for which matrix inversion has to be brought into play.

What is the **Box-Cox transformation**?

The motivation for considering power transformations of our data (y_i) is the frequent occurrence of skewed data, for which a log-transformation, or a reciprocal transformation of (y_i) , will produce a nearly normal distribution, as revealed by the qq-plot, for example.

(picture to be sketched by you).

We follow the notation of Venables and Ripley (1997).

Define $y^{(\lambda)} = (y^\lambda - 1)/\lambda$ for $\lambda \neq 0$,

and define $y^{(\lambda)} = \log(y)$ for $\lambda = 0$.

(You can check that this behaves properly, ie gives a continuous function, as $\lambda \rightarrow 0$.)

We now consider the following generalization of the standard Linear Model:

$$y_i^{(\lambda)} = x_i^T \beta + \epsilon_i$$

for $i = 1, \dots, n$. with $(\epsilon_i) \sim NID(0, \sigma^2)$, and so for each value of λ we maximize the log-likelihood function with respect to β, σ^2 and then consider the resulting maximized likelihood as a function of λ ; this is called the **profile likelihood**.

The formula for this profile log-likelihood function is

$$\hat{L}(\lambda) = \text{constant} - (n/2) \log RSS(z^{(\lambda)})$$

where $z^{(\lambda)} = y^\lambda / y_{gm}^{\lambda-1}$. Here y_{gm} denotes the geometric mean of the observations, so that $\log y_{gm} = (1/n) \sum \log(y_i)$ and

$RSS(z^{(\lambda)})$ is the residual sum of squares in the regression $E(z^{(\lambda)}) = X\beta$.

The Venables and Ripley MASS library function displays this profile log-likelihood

function, together with an approximate 95% likelihood ratio confidence interval for λ (based on the χ^2 distribution with 1 df). For the current example $\lambda = 0$ is the suggested transformation, corresponding to the model

$$\omega_l : E(\log(y_{ij})) = \mu + \alpha_i + \beta_j.$$

Remarks on the derivation of the expression for $\hat{L}(\lambda)$.

First note that if $Z = Y^{(\lambda)}$ is distributed as $N(\mu, \sigma^2)$ so that Z has pdf $g(z)$, given by

$$g(z) \propto (1/\sigma) \exp -(z - \mu)^2 / 2\sigma^2$$

then Y has pdf say $h(y)$, given by

$$h(y) = g(z(y))|J|$$

where $J = \frac{dz}{dy}$, following the usual rule for the pdf of the transform of a random variable.

Hence for the given observations (y_i, x_i) the loglikelihood function is

$$l(\lambda, \beta, \sigma^2) = -(n/2) \log \sigma^2 + (\lambda - 1) \Sigma \log y_i - \frac{\Sigma (y_i^{(\lambda)} - \beta^T x_i)^2}{2\sigma^2}$$

for $\lambda \neq 0$, and for $\lambda = 0$ we find that

$$l(0, \beta, \sigma^2) = -(n/2) \log \sigma^2 - \Sigma \log y_i - \frac{\Sigma (\log y_i - \beta^T x_i)^2}{2\sigma^2}.$$

We can easily maximise each of these two, for fixed λ , with respect to β, σ^2 to obtain the expression for the profile loglikelihood given above.

(In the present example $y_i > 0$ for all i .)

A refinement of the Box-Cox transformation would be to consider replacing y by $y + \alpha$ in the above transformation, where the parameter α which also has to be estimated, may be introduced to avoid difficulties caused by negative values of y . See Venables and Ripley (1997) p216.)

Here are the parameter estimates for the model ω , which corresponds to the untransformed prices.

Call: `lm(formula = p ~ author + country, na.action = na.omit)`

Coefficients:

	Value	Std. Error	t value	Pr(> t)
(Intercept)	12.9210	1.1957	10.8063	0.0000
author2	2.3520	1.4134	1.6641	0.1077
author3	0.0280	1.4134	0.0198	0.9843
author4	-1.0860	1.6607	-0.6539	0.5187
author5	1.8043	1.5102	1.1947	0.2426
author6	-0.1140	1.4134	-0.0807	0.9363
author7	8.1708	1.5116	5.4055	0.0000
author8	0.7357	1.6602	0.4431	0.6612
author9	-2.7340	1.6603	-1.6467	0.1112
author10	-4.7863	1.5115	-3.1665	0.0038
countryGer	1.7730	1.0361	1.7112	0.0985
countryFra	-1.8778	1.1312	-1.6601	0.1085
countryUS	-2.0337	1.0778	-1.8868	0.0700

```
countryAustria  1.1538  1.1308  1.0204  0.3166
```

Residual standard error: 2.235 on 27 degrees of freedom

Multiple R-Squared: 0.803

F-statistic: 8.467 on 13 and 27 degrees of freedom, the p-value is 1.796e-06

The standard diagnostic plots for the above model are less satisfactory than those obtained for the model ω_l below, which corresponds to the additive model for the log-transformed data, or equivalently, to the model

$$\omega_l : p_{ij} = a_i b_j \exp \epsilon_{ij}.$$

```
Call: lm(formula = lp ~ country + author, na.action = na.omit)
```

Coefficients:

	Value	Std. Error	t value	Pr(> t)
(Intercept)	2.5612	0.0703	36.4172	0.0000
countryGer	0.0950	0.0609	1.5597	0.1305
countryFra	-0.1985	0.0665	-2.9832	0.0060
countryUS	-0.1554	0.0634	-2.4512	0.0210
countryAustria	0.0646	0.0665	0.9712	0.3401
author2	0.1855	0.0831	2.2317	0.0341
author3	0.0202	0.0831	0.2429	0.8099
author4	-0.0855	0.0977	-0.8753	0.3891
author5	0.1521	0.0888	1.7127	0.0982
author6	0.0091	0.0831	0.1095	0.9136
author7	0.4602	0.0889	5.1760	0.0000
author8	0.0604	0.0977	0.6186	0.5414
author9	-0.2290	0.0977	-2.3448	0.0266
author10	-0.4888	0.0889	-5.4975	0.0000

Residual standard error: 0.1314 on 27 degrees of freedom

Multiple R-Squared: 0.8665

F-statistic: 13.49 on 13 and 27 degrees of freedom, the p-value is 1.375e-08

Here's a thought for you to consider. This same basic 'unbalanced design' is obtained in a standard examination, where the rows or authors in the table correspond to the different candidates taking this examination, and the columns or countries correspond to the different questions set, which may not be equally difficult. The examiners' task is to rank the candidates.

You will see that if the design becomes sufficiently unbalanced, then it will be impossible to compare certain subsets of candidates. Experiment with the given data set, by making rather more of the p_{ij} as NA entries.

Remark on approximate distribution of transformation of a random variable

Suppose that Y has approximate distribution $N(\mu, \sigma^2)$, and $f(\cdot)$ is a 'well-behaved' transformation. Then, to first order, we may write

$$f(Y) = f(\mu) + (Y - \mu)f'(\mu)$$

by expanding the function around μ . Hence, to first order, $f(Y)$ is approximately a linear function of Y , and thus its approximate distribution is

$$N(f(\mu), \sigma^2(f'(\mu))^2).$$

Experiment

- i) with $Y = X/n$, where X is Binomial, parameters n, p , and $f(Y) = \arcsin(\sqrt{Y})$,
- ii) with $Y = X/n$, where X is Poisson, mean $n\lambda$ and $f(Y) = \sqrt{Y}$.

Notes on Solution to Worksheet 6.

An introduction to binomial regression

Suppose that we have independent observations $Y_i \sim Bi(n_i, \pi_i)$, $1 \leq i \leq n$, thus

$$f(y_i|\pi_i) \propto \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$

and so the resulting loglikelihood is say $L(\pi)$, where

$$\begin{aligned} L(\pi) &= \text{constant} + \Sigma[y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i)] \\ &= \text{constant} + \Sigma[y_i/n_i \log(\pi_i) + (1 - (y_i/n_i)) \log(1 - \pi_i)] \times n_i. \end{aligned}$$

First we fit Ω , the model $0 \leq \pi_i \leq 1$, $1 \leq i \leq n$.

It is easily checked that $L(\pi)$ is maximised subject to π in Ω by

$$\hat{\pi}_i = y_i/n_i, \quad 1 \leq i \leq n,$$

and the resulting maximised loglikelihood is

$$L(\hat{\pi}) = \text{constant} + \Sigma[y_i \log(y_i/n_i) + (n_i - y_i) \log((n_i - y_i)/n_i)].$$

Now we try fitting ω , in which π depends, via a given **link** function $g()$ say, on the given explanatory covariates x_i . Specifically, we try

$$\omega : g(\pi_i) = \beta^T x_i, \quad 1 \leq i \leq n$$

where β is an unknown parameter, of dimension p , say.

Our aim is

i) to assess the fit of ω , via Wilks' theorem,

ii) to produce $\hat{\beta}$, and its associated covariance matrix, and hence to find confidence intervals for the individual components of β , and to see if we can then simplify the model by dropping 'unnecessary' components β_ν , etc.

Note that the link function

$$g(\pi) = \log(\pi/(1 - \pi))$$

(the logit link function) is the *canonical* link function for the binomial. With this particular link function you can see that $\Sigma y_i x_i$ is a sufficient statistic for the parameter β .

The general method of finding $\hat{\beta}$ is to solve

$$\frac{\partial L}{\partial \beta} = 0.$$

This can only be done by iteration. The iteration makes use of

$$-\frac{\partial^2 L}{\partial \beta \partial \beta^T}$$

which is in any case needed, as its value at the mle is the inverse of v , the asymptotic covariance matrix of $\hat{\beta}$.

(Recall that the asymptotic distribution of $\hat{\beta}$ is $N_p(\beta, v)$.)

By Wilks' theorem, we can say that

on ω , the approximate distribution of $2[L(\hat{\pi}) - L(\pi(\hat{\beta}))]$ is χ^2 , with degrees of freedom say f , and $f = n - p$.

We define D as the quantity $2[L(\hat{\pi}) - L(\pi(\hat{\beta}))]$: this is also referred to as 'the residual deviance' under ω . The model ω fits well if D is close to or less than f .

(But, WARNING, we cannot assess the fit of ω this way if Y_i are binary variables, ie if $n_i = 1$ for all i . In this case, although the mle method for estimating β

works as above, Wilks' theorem does not apply to the null distribution of D , so you need to use other methods to assess the fit.)

Let us define π_i^* as the fitted probabilities under ω , so that $g(\pi_i^*) = \hat{\beta}^T x_i$, then

$$D = 2\Sigma(y\log(\hat{\pi}/\pi^*) + (n - y)\log((1 - \hat{\pi})/(1 - \pi^*)))$$

where for convenience we have dropped the suffix i . You can check that D may be rewritten as

$$D = \Sigma[y\log(y/e) + (n - y)\log(n - y)/(n - e)]$$

where $e_i = n_i\pi_i^*$, the 'expected values' under ω .

As usual, D is approximately equal to the corresponding Pearson X^2 , which here is

$$X^2 = \Sigma[(y - e)^2/e + (n - y - (n - e))^2/(n - e)].$$

Thus X^2 is the sum of squares of the **Pearson residuals** $(y - e)/\sqrt{e(n - e)/n}$. The **deviance residuals** are defined in a corresponding way, from the terms in D .

Since $\hat{\beta}$ is approximately distributed as $N_p(\beta, v)$, we can for example test the null hypothesis that $\beta_1 = 0$, by referring $\hat{\beta}_1/\sqrt{v_{11}}$ to the $N(0, 1)$ distribution, and similarly we can compute a 95% confidence interval for β_1 .

Notes on Solution to Worksheet 7.

One student (C.H.Jackson, Trinity Hall) concluded his remarks on the omission of those points for which $n_{fail} = 0$ by the following paragraph:

"It is inconceivable how members of one of the most famous scientific organisations in the world could make such an elementary mistake. Even people with no knowledge of statistics could have realised that the fact that half the flights had no failed rings was important to an analysis of the frequency of ring failure. (*although this tale could possibly be a statistical urban myth.*)"

Here is the suggested program, with comments and one graph.

```
>orings = read.table("orings", header=T)
>six = rep(6, times=23); attach(orings)
> pfail=nfail/six
>first.glm = glm(pfail ~ temp, family = binomial, weights = six)
>summary(first.glm) # we edit the output somewhat
```

```
Call: glm(formula = pfail ~ temp, family = binomial, weights = six)
```

```
Deviance Residuals:
```

```
      Min       1Q   Median       3Q      Max
-0.9522692 -0.7829873 -0.541173 -0.04378912  2.651517
```

```
Coefficients:
```

```
              Value Std. Error  t value
(Intercept)  5.0849772  3.05247412  1.665854
temp        -0.1156012  0.04702362 -2.458364
```

```
(Dispersion Parameter for Binomial family taken to be 1 )
```

```
Null Deviance: 24.23036 on 22 degrees of freedom
```

```
Residual Deviance: 18.08633 on 21 degrees of freedom
```

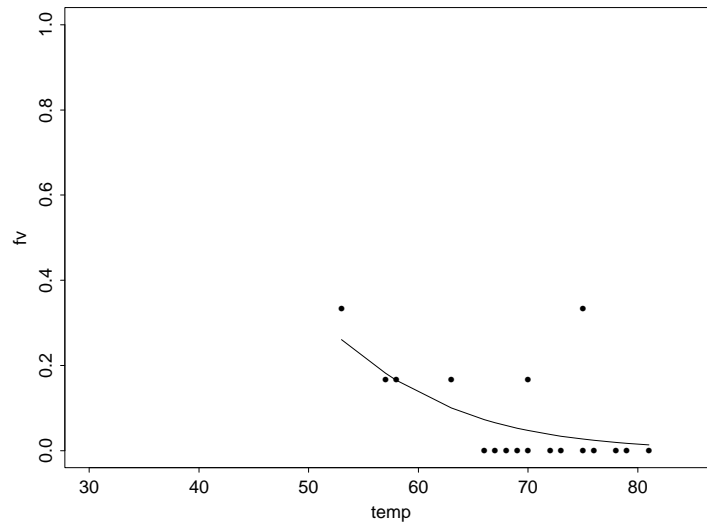


Figure 3: Fitted and observed probabilities of failure for the O-rings data

Number of Fisher Scoring Iterations: 5

Correlation of Coefficients:

(Intercept)

temp -0.993214

This model fits quite well. (Compare the Residual Deviance with its degrees of freedom.) You will also see that the probability of failure **increases** as the temperature reduces, but note that we have no data for a temperature cooler than 53degF. We plot the fitted values, with the observed points superimposed.

```
> fv = first.glm$fitted.values
> plot(temp, fv, type="l", xlim=c(30,85), ylim=c(0,1))
> points(temp, pfail)
```

Can you see the reason for this choice of xlim, ylim?

Now to repeat the logistic regression for the censored data.

```
> wrong.glm = glm(pfail~temp, binomial, weights=six,subset=(nfail >0))
> summary(wrong.glm)
```

Deviance Residuals:

1	2	3	4	13	14	18
0.6890993	-0.2835787	-0.2849574	-0.2918573	-0.3015338	-0.3015338	0.6560362

Coefficients:

	Value	Std. Error	t value
(Intercept)	-1.389527689	3.19575244	-0.43480455
temp	0.001415884	0.04977301	0.02844682

(Dispersion Parameter for Binomial family taken to be 1)

Null Deviance: 1.334694 on 6 degrees of freedom

Residual Deviance: 1.333885 on 5 degrees of freedom

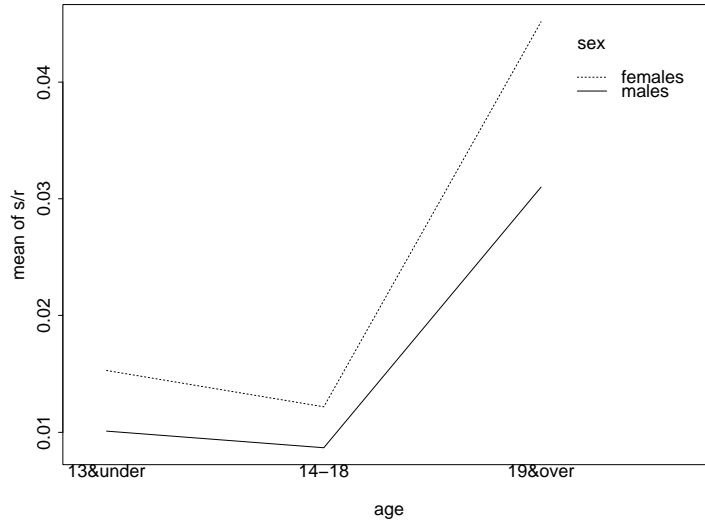


Figure 4: Interaction plot for missing persons

Number of Fisher Scoring Iterations: 4

Correlation of Coefficients:

(Intercept)

temp -0.9930524

Removing the points for which (nfail=0) gives us a model which fits well, but suggests that the probability of failure **does not depend on** the temperature. Alas.

Notes on Solution to Worksheet 8.

We take as the model

$s_{ij} \sim$ independent Binomial, parameters (r_{ij}, π_{ij}) for $1 \leq i \leq 2, 1 \leq j \leq 3$, so that i, j correspond to the factors sex, age, respectively.

r_{ij} = number reported missing during the year

s_{ij} = number still missing at the end of that year, in category (i, j) .

We fit

$$\log \frac{\pi_{ij}}{(1 - \pi_{ij})} = \mu + \alpha_i + \beta_j$$

for $i = 1, 2$ and $j = 1, 2, 3$ with (as usual) $\alpha_1 = \beta_1 = 0$.

You should find that this model fits rather well, which is perhaps a little surprising considering that these are just numbers taken out of the newspaper. The interpretation of the parameter estimates is of course very important.

Here is our interaction plot. Now, with this model, $E(s_{ij}) = r_{ij}\pi_{ij}$, and evidently r is large, π is small.

So for comparison we try the model

s_{ij} independent \sim Poisson, with mean $(r_{ij}\mu_{ij})$,

so that $\log E(s_{ij}) = \log \mu_{ij} + \log r_{ij}$, and we fit

$$\log \mu_{ij} = \mu + \alpha_i + \beta_j.$$

The $\log r_{ij}$ term then enters the glm() as an “offset”.

You should find that you get very similar results to those obtained with the binomial distribution.

Notes on solution to Worksheet 9.

In this worksheet, our aim is to test whether the *deviant behaviour* is significantly associated with the emotionality level. A glance at the data indicates that this must be the case, but for illustrative purposes we proceed to produce various, equivalent, test statistics.

We have data $(a_i, b_i), 1 \leq i \leq 4$.

i) We let $t_i = a_i + b_i$ and we assume that $a_i|t_i$ are independent Binomial, parameters (t_i, π_i) , ie $\text{Bi}(t_i, \pi_i)$. Now

`glm(cbind(a,b) ~ one, binomial)`

fits the model

$$\log \frac{\pi_i}{(1 - \pi_i)} = \mu, 1 \leq i \leq 4,$$

equivalently, the model

$P(\text{subject exhibits deviant behaviour} \mid \text{emotionality level is } i)$ is the same for all i ,

in other words, the behaviour is independent of the emotionality level.

You should find that this model fails to fit; its deviance is 43.21, with 3 df.

Essentially the test statistic will be

$$D = 2 \sum \sum x_{ij} \log(x_{ij}/e_{ij})$$

to be referred to χ^2_3 , where (x_{ij}) is the matrix whose first column is a , and second column is b , and (e_{ij}) are defined below.

ii) Now we use the given S-Plus function to do the Pearson χ^2 test.

In the new notation of (x_{ij}) , we are implicitly assuming that $(x_{ij})|x_{i+}$ are independently distributed, as Multinomial, with parameters $(x_{i+}, (\pi_{ij}))$, say, with $\sum_j \pi_{ij} = 1$ for each i , and the row totals $\sum_j x_{ij}$ are fixed as x_{i+} , for $i = 1, \dots, 4$.

We test $H_0 : \pi_{ij} = \pi_j$ for each i, j , where (π_j) are unknown and add to 1.

The Pearson χ^2 statistic is of course

$$X^2 = \sum \sum (x_{ij} - e_{ij})^2 / e_{ij}$$

where $e_{ij} = (x_{i+}x_{+j})/x_{++}$ for all i, j , the usual “expected values under H_0 ”.

You will see that D, X^2 are numerically very close.

iii) Now we use the Poisson ‘family’ and exploit the equivalence between the Poisson and the multinomial, for loglinear modelling. We assume a new model: x_{ij} are independent Poisson, with means μ_{ij} , and we fit

$$H_0 : \log(\mu_{ij}) = \mu + \alpha_i + \beta_j, 1 \leq i \leq 4, 1 \leq j \leq 2$$

(and `glm()` takes $\alpha_1 = 0, \beta_1 = 0$ as usual).

Our alternative hypothesis is

$$H : \mu_{ij} \geq 0, 1 \leq i \leq 4, 1 \leq j \leq 2$$

(ie that the μ ’s are any positive numbers).

For this last `glm()`, we create the vector y , with 8 elements, consisting of (51, 69, ..., 13). and we also create the two new vectors RR, CC of the same length, which are to be *factors*, indicating the corresponding row and column labels.

You will see that, once again, the deviance for testing the fit of H_0 is

$$D = 2 \sum \sum x_{ij} \log(x_{ij}/e_{ij})$$

Table 1: A 2×2 contingency table

$a_1 = 28$	$b_1 = 22$
$a_2 = 3$	$b_2 = 13$

to be referred to χ_3^2 , a high value showing that H_0 is to be rejected.

iv) Now all of the three methods above use Wilks' theorem, relying on an *asymptotic* approximation to derive the appropriate p-value. Thus now we describe the Fisher exact test for a 2×2 contingency table; S-plus allows us to extend the argument given below to an $r \times c$ table, provided that the sum of the marginal totals is not too big.

Suppose our data matrix is as given in Table 1. Put $t_1 = a_1 + b_1, t_2 = a_2 + b_2$, thus defining the row totals. Assume that, conditional on t_1, t_2 the frequencies a_1, a_2 are independent binomial, with parameters $(t_1, \pi_1), (t_2, \pi_2)$ respectively. We wish to test

$$H_0 : \pi_1 = \pi_2 \text{ against } H_1 : \pi_1 > \pi_2, \text{ say.}$$

Expressed loosely, in terms of the data given in the example, we wish to test whether 28/50 is 'significantly larger than' 3/16. Note that both π_1, π_2 are unknown parameters.

Reparametrise by

$$\log(\pi_1/(1 - \pi_1)) = \mu + \lambda, \quad \log(\pi_2/(1 - \pi_2)) = \mu - \lambda.$$

Then you will see that we wish to test $H_0 : \lambda = 0$ against $H_1 : \lambda > 0$, with μ being a nuisance parameter. Now you can check that the joint frequency function of a_1, a_2 can be written as

$$p(a_1, a_2 | \mu, \lambda) \propto \binom{t_1}{a_1} \binom{t_2}{a_2} e^{(\mu+\lambda)a_1} e^{(\mu-\lambda)a_2}$$

thus

$$p(a_1, a_2 | \mu, \lambda) \propto \binom{t_1}{a_1} \binom{t_2}{a_2} e^{\lambda(a_1 - a_2)} e^{\mu(a_1 + a_2)}.$$

Hence it can be shown that the distribution of $(a_1 - a_2)$, conditional on the column total $a_1 + a_2$, depends on λ, μ only through λ .

Hence we base our test on the distribution of $a_1 | a_1 + a_2 = a$ say, a being the observed total. (This is a special – and very important case – of a test being constructed by conditioning on an *ancillary* statistic, here $a_1 + a_2$.)

With this conditional test, we reject H_0 in favour of H_1 if

$$a_1 > C$$

where C is defined by

$$P_{H_0}(A_1 > C | A_1 + A_2 = a) = \alpha$$

α being the size of the test, for example 0.05. Now we can easily see that, on H_0 , the distribution of $A_1 | A_1 + A_2 = a$ is *hypergeometric*, that is,

$$P_{H_0}(A_1 = a_1 | A_1 + A_2 = a) = \frac{P_{H_0}(A_1 = a_1, A_2 = a - a_1)}{P_{H_0}(A_1 + A_2 = a)} = \frac{\binom{t_1}{a_1} \binom{t_2}{a_2}}{\binom{t_1 + t_2}{a}}$$

since, on $H_0, A_1 \sim Bi(t_1, \pi), A_2 \sim Bi(t_2, \pi)$, independent, defining π in the obvious way by $\log(\pi/(1 - \pi)) = \mu$.

Thus the 'exact' significance level is found from the tail of this distribution, and

for the general $r \times c$ table this argument is extended to give the *multivariate hypergeometric* distribution.

Notes on Solution to Worksheet 13.

Here we have data (n_{ijkl}) for $i, j, k, l = 1, 2$ and where the subscripts i, j, k, l correspond, respectively to

sex ('gender' is the politically correct term), depression, behavioural symptoms, and anxiety symptoms.

We read in (n_{ijkl}) into a vector of length 16 and set up the nested design into 4 vectors, also each of length 16. The command

```
glm.sat = glm(n~anx*beh*dep*sex,poisson)
```

has the effect of fitting the 'saturated' model

$$\Omega : (n_{ijkl}) \sim Mn(n, (p_{ijkl}))$$

where $\sum_{ijkl} p_{ijkl} = 1$.

This will give us a perfect fit, ie deviance= 0 and df = 0, but it is a good place to start, since we can glance down the parameter estimates and their se's to get some idea of which high order interaction terms we can drop. (The syntax of glm() will not allow us to do something silly, such as dropping a 2-way interaction, say $dep \times anx$ before dropping the 3-way interaction $dep \times anx \times beh$.)

Our general strategy in model-fitting will be to start with a very full model (which will fit perfectly but will almost certainly fail to help us interpret the data at all) and then drop high order interaction terms, until we reach the simplest possible model consistent with the data. There may be more than one candidate for this special position, if the data-set is quite complex.

```
gl.three = glm(n ~ (anx + beh + dep + sex)^3,poisson)
```

has the effect of fitting

$$H_3 : \log(p_{ijkl}) = \mu + \alpha_i + \dots + \alpha_{ijkl}$$

(ie including all terms except the 4th-order effect).

This model fits well, but is still too complex to be helpful.

In the current example, there is a total of $(1 + 1 + 4 + \dots + 1)$ possible log-linear models, ranging from the saturated model Ω to the model of complete independence

$$\log(p_{ijkl}) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l$$

to the 'null' model

$$\log(p_{ijkl}) = \mu.$$

(It is a good exercise for you to enumerate all the possible models.)

In general it would be tedious, and quite possibly silly, to check the fit of *all* possible loglinear models. This is where the

```
step.glm()
```

function will prove very helpful. You will see that this function (in the way we use it here) steps *down* from the saturated model, checking which parameters it is safe to drop. The criterion used for deciding which parameters to drop is the

Akaike Information Criterion.

See Venables and Ripley (2nd ed, p221) for the definition and use of the *AIC*. It is defined as

$AIC = -2 \text{ maximized log-likelihood} + 2 \text{ number of parameters in the model.}$

So you see that in adopting the *AIC* as our model-fitting criterion, we do not simply go for the model with the highest maximized log-likelihood, which would automatically be the model with the most parameters (assuming that we are comparing *nested* models) but we use the *AIC* to incorporate a ‘trade-off’ between the log-likelihood and the number of parameters fitted.

Caution:

See also Venables and Ripley p227 for the delicate question of the *AIC* for a glm with unknown scale parameter ϕ . (Not needed for the current example, where $\phi = 1$, but you may encounter this problem later, so you should be aware of it.)