

Preliminaries

```
> library(DAAG)
```

Exercise 1

The following table shows numbers of occasions when inhibition (i.e., no flow of current across a membrane) occurred within 120 s, for different concentrations of the protein peptide-C (data are used with the permission of Claudia Haarmann, who obtained these data in the course of her PhD research). The outcome *yes* implies that inhibition has occurred.

conc	0.1	0.5	1	10	20	30	50	70	80	100	150
no	7	1	10	9	2	9	13	1	1	4	3
yes	0	0	3	4	0	6	7	0	0	1	7

Use logistic regression to model the probability of inhibition as a function of protein concentration.

It is useful to begin by plotting the logit of the observed proportions against $\log(\text{conc})$. Concentrations are nearer to equally spaced on a scale of relative dose, rather than on a scale of dose, suggesting that it might be appropriate to work with $\log(\text{conc})$. In order to allow plotting of cases where *no* = 0 or *yes* = 0, we add 0.5 to each count.

```
> conc <- c(0.1, 0.5, 1, 10, 20, 30, 50, 70, 80, 100, 150)
> no <- c(7, 1, 10, 9, 2, 9, 13, 1, 1, 4, 3)
> yes <- c(0, 0, 3, 4, 0, 6, 7, 0, 0, 1, 7)
> n <- no + yes
> plot(log(conc), log((yes + 0.5)/(no + 0.5)))
```

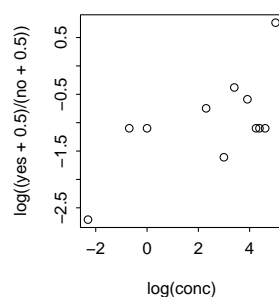


Figure 1: Plot of $\log((\text{yes}+0.5)/(\text{no}+0.5))$, against $\log(\text{conc})$.

The plot seems reasonably consistent with the use of $\log(\text{conc})$ as the explanatory variable.

The code for the regression is:

```
> p <- yes/n
> inhibit.glm <- glm(p ~ I(log(conc)), family = binomial, weights = n)
> summary(inhibit.glm)
```

2

Call:

```
glm(formula = p ~ I(log(conc)), family = binomial, weights = n)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.251	-1.060	-0.503	0.315	1.351

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.766	0.521	-3.39	0.0007
I(log(conc))	0.344	0.144	2.39	0.0170

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 16.6834 on 10 degrees of freedom
Residual deviance: 9.3947 on 9 degrees of freedom
AIC: 29.99

Number of Fisher Scoring iterations: 4

Exercise 2

In the data set (an artificial one of 3121 patients, that is similar to a subset of the data analyzed in Stiell et al. (2001)) `minor.head.injury`, obtain a logistic regression model relating `clinically.important.brain.injury` to other variables. Patients whose risk is sufficiently high will be sent for CT (computed tomography). Using a risk threshold of 0.025 (2.5%), turn the result into a decision rule for use of CT.

```
> sapply(head.injury, range)
```

```
      age.65 amnesia.before basal.skull.fracture GCS.decrease GCS.13
[1,]      0          0                0          0          0
[2,]      1          1                1          1          1
      GCS.15.2hours high.risk loss.of.consciousness open.skull.fracture vomiting
[1,]          0          0                0          0          0
[2,]          1          1                1          1          1
      clinically.important.brain.injury
[1,]          0
[2,]          1
```

```
> injury.glm <- glm(clinically.important.brain.injury ~ ., data = head.injury,
+ family = binomial)
> summary(injury.glm)
```

Call:

```
glm(formula = clinically.important.brain.injury ~ ., family = binomial,
     data = head.injury)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.277	-0.351	-0.210	-0.149	3.003

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.497	0.163	-27.61	< 2e-16
age.65	1.373	0.183	7.52	5.6e-14
amnesia.before	0.689	0.172	4.00	6.4e-05
basal.skull.fracture	1.962	0.206	9.50	< 2e-16
GCS.decrease	-0.269	0.368	-0.73	0.46515
GCS.13	1.061	0.282	3.76	0.00017
GCS.15.2hours	1.941	0.166	11.67	< 2e-16
high.risk	1.111	0.159	6.98	2.9e-12
loss.of.consciousness	0.955	0.196	4.88	1.1e-06
open.skull.fracture	0.630	0.315	2.00	0.04542
vomiting	1.233	0.196	6.29	3.2e-10

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1741.6 on 3120 degrees of freedom
 Residual deviance: 1201.3 on 3110 degrees of freedom
 AIC: 1223

Number of Fisher Scoring iterations: 6

Observe that $\log(.025/(1-.025)) = -3.66$, an increase of 0.84 above the intercept (= -4.50). This change in risk results from (1) GCS.decrease with any other individual factor except amnesia.before, GCS.decrease and open.skull.fracture; (2) GCS.decrease with any two of amnesia.before, open.skull.fracture and loss.of.consciousness; (3) any of the individual factors age.65, basal.skull.fracture, GCS.15.2hours, high.risk and vomiting, irrespective of the levels of other factors.

Exercise 3

Consider again the moths data set of Section 8.4.

- (a) What happens to the standard error estimates when the `poisson` family is used in `glm()` instead of the `quasipoisson` family?
- (b) Analyze the P moths, in the same way as the A moths were analyzed. Comment on the effect of transect length.

- (a) The dispersion estimate was 2.69. Use of the `quasipoisson` family has the effect of increasing SEs by a factor of $\sqrt{2.69}$, relative to the `poisson` family. See the first two lines on p.215. SEs on pp.214-215 will thus be reduced by this factor if the `poisson` family is (inappropriately) specified.

- (b) `> sapply(split(moths$P, moths$habitat), sum)`

Lowerside	Bank	Disturbed	NEsoak	NWsoak	SEsoak	SWsoak	Upperside
17	4	33	14	19	6	48	8

```
> moths$habitat <- relevel(moths$habitat, ref = "Lowerside")
> P.glm <- glm(P ~ habitat + log(meters), family = quasipoisson,
+ data = moths)
```

The highest numbers are now for SWsoak and for Disturbed. The number of moths increases with transect length, by a factor of approximately 1.74 (= $e^{.55}$) for each one meter increase in transect length.

*Exercise 4**

The factor `dead` in the data set `mifem` (*DAAG* package) gives the mortality outcomes (`live` or `dead`), for 1295 female subjects who suffered a myocardial infarction. (See Section 11.5 for further details.) Determine ranges for `age` and `yr onset` (year of onset), and determine tables of counts for each separate factor. Decide how to handle cases for which the outcome, for one or more factors, is not known. Fit a logistic regression model, beginning by comparing the model that includes all two-factor interactions with the model that has main effects only.

First, examine various summary information:

```
> str(mifem)

'data.frame':      1295 obs. of  10 variables:
 $ outcome : Factor w/ 2 levels "live","dead": 1 1 1 1 2 1 1 2 2 2 ...
 $ age      : num  63 55 68 64 67 66 63 68 46 66 ...
 $ yr onset : num  85 85 85 85 85 85 85 85 85 85 ...
 $ premi    : Factor w/ 3 levels "n","y","nk": 1 1 2 1 1 1 1 2 1 2 ...
 $ smstat   : Factor w/ 4 levels "n","c","x","nk": 3 2 4 3 4 3 1 1 2 2 ...
 $ diabetes: Factor w/ 3 levels "n","y","nk": 1 1 3 1 3 3 1 1 1 1 ...
 $ highbp   : Factor w/ 3 levels "n","y","nk": 2 2 2 2 3 3 2 2 2 2 ...
 $ hichol   : Factor w/ 3 levels "n","y","nk": 2 2 3 1 3 3 1 2 3 1 ...
 $ angina   : Factor w/ 3 levels "n","y","nk": 1 1 2 2 3 3 1 2 3 1 ...
 $ stroke   : Factor w/ 3 levels "n","y","nk": 1 1 1 1 3 3 1 2 1 2 ...

> sapply(mifem[, c("age", "yr onset")], range)

      age yr onset
[1,]  35     85
[2,]  69     93

> lapply(mifem[, -(1:3)], table)

$premi

  n  y nk
928 311 56

$smstat

  n  c  x nk
522 390 280 103

$diabetes

  n  y nk
978 248 69

$highbp

  n  y nk
406 813 76
```

```
$hichol
```

```
  n   y  nk
655 452 188
```

```
$angina
```

```
  n   y  nk
724 472  99
```

```
$stroke
```

```
  n   y  nk
1063 153  79
```

For all of the factors, there are a large number of nk's, i.e., *not known*. A straightforward way to handle them is to treat nk as a factor level that, as for y and n, may give information that helps predict the outcome. For ease of interpretation we will make n, the reference level.

```
> for (j in 4:10) mifem[, j] <- relevel(mifem[, j], ref = "n")
> mifem1.glm <- glm(outcome ~ ., family = binomial, data = mifem)
> mifem2.glm <- glm(outcome ~ .^2, family = binomial, data = mifem)
> anova(mifem1.glm, mifem2.glm)
```

Analysis of Deviance Table

```
Model 1: outcome ~ age + yronset + premi + smstat + diabetes + highbp +
  hichol + angina + stroke
```

```
Model 2: outcome ~ (age + yronset + premi + smstat + diabetes + highbp +
  hichol + angina + stroke)^2
```

	Resid. Df	Resid. Dev	Df	Deviance
1	1277	1173		
2	1152	1014	125	159

```
> CVbinary(mifem1.glm)
```

```
Fold:  2 10 6 8 9 7 1 4 3 5
```

```
Internal estimate of accuracy = 0.807
```

```
Cross-validation estimate of accuracy = 0.804
```

```
> CVbinary(mifem2.glm)
```

```
Fold:  3 4 9 6 7 1 10 5 8 2
```

```
Internal estimate of accuracy = 0.839
```

```
Cross-validation estimate of accuracy = 0.781
```

The difference in deviance seems statistically significant ($\text{pchisq}(125,159) = 0.021$), but it may be unwise to trust the chi-squared approximation to the change in deviance.

It is safer to compare the cross-validated accuracy estimates, which in individual cross-validation runs were marginally lower for `mifem2.glm` than for `mifem1.glm`; 0.78 as against 0.80. Note also that there were convergence problems for the model that included all first order interaction terms.