Statistical Modeling, 2005, Solutions to Ex 3

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

January 20, 2005

1. Clearly $log(f(t_i|\alpha, \beta)) = -\nu log(\mu_i) - t_i/\mu_i +$ constant .
Hence loglikelihood is

$$L(\alpha, \beta) = \Sigma \nu log(\alpha + \beta x_i) - \alpha \Sigma t_i - \beta \Sigma x_i t_i + constant$$

where Σ runs from $i = 1$ to $i = n$.
Hence, by the factorisation theorem, (a_1, a_2) is sufficient for (α, β).
Take $x_i = 0$ for i=1,...,m, and x_i=1 for $i = m + 1, ..., n$.
Then

$$L(\alpha, \beta) = \nu m log(\alpha) + \nu(n - m) log(\alpha + \beta) - \alpha(T_1 + T_2) - \beta T_2 + constant$$

where $T_1 = t_1 + ... + t_m, T_2 = t_{m+1} + ... + t_n$.
To estimate β, solve $\partial L/\partial\alpha = 0, \partial L/\ \partial\beta = 0$ for $(\hat{\alpha}, \hat{\beta})$.
We know that the asymptotic distribution of $(\hat{\alpha}, \hat{\beta})$ is bivariate normal, with mean (α, β), and covariance matrix V, say, where $-V^{-1}$ is the expectation of the matrix of 2nd partial derivatives of L.
Now

$$\partial L/\partial\alpha = (\nu m/\alpha) + \nu(n - m)/(\alpha + \beta) - (T_1 + T_2)$$

$$\partial L/\partial\beta = \nu(n - m)/(\alpha + \beta) + T_2$$

Hence, after evaluating the matrix of 2nd derivatives of L, (whose elements turn out not to be random variables), we see that V^{-1} is of the form

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

(where you will find the values of a, b, c).
Hence, inverting V shows us that $\hat{\beta}$ is approx'ly $N(\beta, a/(ac - b^2))$.
Thus the approximate confidence interval for β will be $\hat{\beta} \pm 2\sqrt{(}a/(ac - b^2))$.

2. With the given model, we see that the mle's of (α, β) minimise

$$\Sigma\Sigma(y_{ij} - \alpha - \beta x_{ij})^2$$

where in $\Sigma\Sigma$, $\ j = 1, ..., n_i, i = 1, ..., t$.
But $x_{ij} = x_i$, for all i, j so that the mle's of (α, β) minimise

$$\Sigma\Sigma(y_{ij} - \alpha - \beta x_i)^2$$

giving $\hat{\beta} = S_{xy}/Sxx$ where $S_{xy} = \Sigma\Sigma(y_{ij} - \bar{y})x_i$ and S_{xx} is defined similarly.
Now it is true that

$$\bar{y}_i = \alpha + \beta x_i + \eta_i \text{ say}$$

where η_i is dist'd as $NID(0, \sigma^2/n_i)$.
If we choose (α, β) to minimise $\Sigma(\bar{y}_i - \alpha - \beta x_i)^2$, we will obtain less accurate estimators of (α, β) than $(\hat{\alpha}, \hat{\beta})$. You can check that the estimators thus obtained will still be unbiased, but will have larger

variances.
However, if we choose (α, β) to minimise

$$\Sigma(\bar{y}_i - \alpha - \beta x_i)^2 n_i$$

then we will obtain $(\hat{\alpha}, \hat{\beta})$ as above, ie the correct estimators. So using only the reduced data set $(x_i, \bar{y}_i, n_i)$ in this way will be fine, as long as σ^2 is known.
If σ^2 is unknown, as is almost always the case in practice, then we will get a more accurate estimator of it by using the ORIGINAL data set(and estimating it as usual by $(residual\ ss)/df$) than by using the corresponding expression when we have "condensed" the data down to $(x_i, \bar{y}_i, n_i)$, namely

$$\frac{1}{(t-2)}\Sigma(\bar{y}_i - \hat{\alpha} - \hat{\beta}x_i)^2 n_i.$$

Furthermore, if we only have the "condensed" dataset, we are much less likely to be able to do an adequate job of checking the validity of the original linear model with constant error variance (Sketch some simple examples with $t = 3$).

3. Write Y_{ij} for the binary observation. We will assume that Y_{ij} dist'd independently as $Bi(1, p(x_i))$ for $j = 1, ..., n_i, i = 1, ..., t$.
Thus the loglikelihood of the data is

$$L(\alpha, \beta) = \Sigma\Sigma[y_{ij} log p(x_i) + (1 - y_{ij}) log(1 - p(x_i))]$$

where $log(p(x_i)/(1 - p(x_i)) = logit(p(x_i)) = \alpha + \beta x_i$. Hence

$$L(\alpha, \beta) = \Sigma(y_{i+} log(p(x_i)) + (n_i - y_{i+}) log(1 - p(x_i)).$$

Clearly, the loglikelihood is the same whether we enter the data as
$(0, x_1), (1, x_1), (0, x_1), ...$ ie 15 separate cases,
or as $(1, 3, x_1), ...$ ie 4 separate cases.
Thus, since $(\hat{\alpha}, \hat{\beta})$, and its asymptotic covariance matrix, are obtained purely from the loglikelihood function, we will get the same answers for these whichever of the 2 ways we choose to enter the data. (But try a simple numerical example. Why do you get different expressions for the deviances and their df's ?)

4.a) Minimising $(Y - X\beta)^T(Y - X\beta)$ in β gives

$$(X^T X)\hat{\beta} = X^T Y$$

Hence, if X is of full rank, $\hat{\beta} = (X^T X)^{-1} X^T Y$ so that
b) $\hat{Y} = X\hat{\beta} = HY$ say, where
$H = X(X^T X)^{-1} X^T$.
It is easy to check that $H = H^T$ and $HH = H$,(ie H is a projection matrix).
c) $e = Y - X\hat{\beta} = (I - H)Y = (I - H)\epsilon$.
Now, ϵ dist'd as $N(0, \sigma^2 I)$. Thus, using $(I - H)(I - H)^T = (I - H)$ we see that e is dist'd as $N(0, \sigma^2(I - H))$
giving e_i as $N(0, \sigma^2(1 - h_i))$ as required.
d) Suppose u is an eigen vector of H corresponding to eigen value λ .
Then $Hu = \lambda u$, thus $HHu = \lambda Hu$ thus $Hu = \lambda Hu$ thus
$Hu = 0$ or $\lambda = 1$,so that $\lambda = 0$ *or* 1.
Hence $h_1 + ... + h_n = trace(H) = \lambda_1 + \lambda_2 + ... + \lambda_n$ where (λ_i) are the eigen values of H. Hence

$$(h_1 + ... + h_n) = rank(H) = rank(X) = p.$$

e) Clearly, $var(e_i) = (1 - h_i)\sigma^2$, hence $h_i \leq 1$.
Further, $\hat{Y} = HY = H(X\beta + \epsilon)$,
hence $var(\hat{Y}_i) = h_i \sigma^2$ hence $h_i \geq 0$.

For the case of simple linear regression, the first column of X is 1 (the vector of 1's) and the second column of X is say x', where $x_i' = x_i - \bar{x}$.
Hence $X^T X$ is $diag(n, S_{xx})$ where $S_{xx} = \Sigma(x_i - \bar{x})^2$.
Evaluating H from this shows that

$$h_i = 1/n + (x_i - \bar{x})^2 / S_{xx}$$

f) We see from e) above that x_i distant from $\bar{x}$ will give rise to relatively large h_i.(Sketch graphs of various possible configurations to get a feel for (h_i)) .

5. The solution to this is in the lecture notes.
The practical importance is of course that there is no need for a *multinomial* distribution within glm(): so long as we are fitting log-linear models we can use the Poisson distribution.

6. This example arose from some psychiatry data provided by Professor I.Goodyer of Cambridge University. The 4 factors were
$S = 1, 2$ for girls,boys
$D = 1, 2$ for depression no or yes
$A = 1, 2$ for anxiety symptoms no or yes
$B = 1, 2$ for behavioural symptoms no or yes.
Thus the table is in fact rather *sparse* for the large sample theory to be realistic, but we give this analysis as an illustrative example of the glm( ) modelling.

$$(S) \quad log(p_{ijkl}) = \mu + \alpha_i + ... + (\alpha\beta)_{ij} + ... + (\alpha\beta\gamma)_{ijk} + ... + (\alpha\beta\gamma\delta)_{ijkl}$$

is the saturated model (we assume the usual glm() constraints $\alpha_1 = 0$ etc.
a) fits the saturated model (ie 2^4 parameters for 2^4 observations)so we expect to get deviance 0 with df = 0.
However a quick glance at the parameter estimates and their se's for this model will usually suggest to us which of the high-order interactions(if any) can be dropped.Our object is to fit the simplest possible model which is consistent with the data, and of course we want to interpret this model to the scientist who provided the data.
b) fits (S) with

$$H_0 : (\alpha\beta\gamma\delta)_{ijkl} = 0$$

ie no 4th order interaction. Refer the increase in deviance (2.72) to χ^2_1 to test H_0 (applying Wilks' theorem). The result is non-significant, so we accept H_0. c) fits (S) with H_0 and $H_1 : (\alpha\gamma\delta)_{ikl} = 0, (\beta\gamma\delta)_{jkl} = 0$. Refer the resulting increase in deviance (ie 3.42 - 2.72) to χ^2_2 to test H_1 assuming H_0 true. You find that you accept H_1.
d) fits

$$H_2 : log(p_{ijkl}) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\beta\delta)_{jl}$$

so we have dropped yet more parameters in moving from c) to d). The increase in deviance is (8.48 - 3.42) which is non-significant when referred to χ^2_5 so dropping these extra parameters was permissible. Furthermore, we can assess the fit of the model H_2 by referring 8.48 to χ^2_8. In fact 8.48 is almost the same as $E(\chi^2_8)$, so you see that H_2 is a good fit. You could check that no further parameters can safely be dropped.
The final model is therefore $p_{ijkl} = a_{ij} b_{jk} c_{jl}$ for some a, b, c. We will rewrite this in a more enlightening way as

$$Pr(S = i, B = k, A = l | D = j) = Pr(S = i | D = j) Pr(B = k | D = j) Pr(A = l | D = j)$$

ie conditional on D, the 3 variables S, B, A are independent.
A more helpful way of expressing this (since S clearly is a different sort of variable from D, B and A) is to say
D depends directly on S, but B and A only depend on S *via* D, and given D, the variables B and A

are independent. This is a simple example of a *graphical* model of conditional independence. Draw a graph in which the 4 points S,D,B,A are connected only by the arcs SD,DB and DA.
Finally, the model in which S,D,B,A are independent corresponds to a graph with no links at all between the 4 points, and in glm terms it is

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

7. a) As with most regression problems,it is best to START by doing some simple plots: in this case all possible pairwise plots of the 4 variable concerned.This reveals that price is positively related to each of og, percent and calories, but also, as a student doing this problem in an examination wrote,
"as any experienced lager-drinker knows", the 3 variables og, percent and calories are all measuring more or less the same thing, so we could not reasonably expect 2 or 3 of them to give a much better prediction of price than just one of them.(Were teetotal students at a disadvantage in this examination? Setting real problems in statistics examinations is fraught with hazards for the setters in these politically correct times. Any question which mentions differences between males and females, educational levels, any experiment involving animals ... and so forth is liable to provoke protest by some aggrieved parties)
Linear regression of price on these 3 "explanatory" variables bears out this observation. In fact, og is the best single predictor. Inspection of the residuals reveals that some lagers are priced at a ridiculously high level, but perhaps they are connoisseurs' lagers, ie not aimed at those interested only in the alcohol content.
b) This question is asking the student to set up a 2-level factor (tasted/not tasted) and to do the regression of price on og, percent, calories and this new 2-level factor.
The SPlus6 output and corresponding graph are given at the end of these solutions.

8. Setting up the 3 factors rmq, mcr, events, each with 2 levels (level 2 being the "bad" one) with binomial regression of case on these 3, and the logistic link function, and "total" as the binomial denominator, gives

$$logP(case)/P(control) = -1.248(.253)+1.716(.514)rmq(2)+1.257(.561)mcr(2)+1.620(.330)events(2)$$

with deviance 5.053, df=4.
Thus this model fits well, and none of the 3 factors can be dropped from the model. (Either compare each estimate with its se, or compare the increase in deviance with χ^2_1 when each factor is dropped from the model in turn).
Note: having rmq at level 2 rather than level 1 increases the odds in favour of being a case by a factor of exp(1.716). Having all of rmq,mcr,events at the "bad" level rather than the "good"level increases the odds of being a case by a factor of $\exp(1.716 + 1.257 + 1.620)$. [We could use the covariance matrix of these estimates to find the corresponding confidence interval for this odds ratio.]
Table 2 is analysed similarly.

9. Let M be the Mother's educational level and F the Father's. Under ω,
$(M, F) = (M_1, F_1)$ with probability θ, where $M_1 = F_1$ and $P(M_1 = i) = \phi_i$.
$(M, F) = (M_2, F_2)$ with probability $1-\theta$, where M_2, F_2 are independent, with $P(M_2 = i) = \alpha_i, P(F_2 = j) = \beta_j$,
ie, with probability θ, the mother's education MUST be identical to the father's,
and with probability $1 - \theta$, they are independent (in which case there is a chance that they are identical).
It is easily seen that ω is equivalent to the model
$\log p_{ij} = a_i + b_j$ for $i \neq j$.
If we set up ma, pa as the 4-level factors corresponding respectively to mother's, father's educational level, then using the Poisson family and log link function the glm of n(the frequency) on $(ma + pa)$, gives a deviance of 159.25,which is clearly hugely significant compared with χ^2_9. This tests the hypothesis of *independence* of the factors ma, pa , so common-sense suggests that this hypothesis is unlikely

to hold: in any case, a glance at the 4×4 contingency table shows that the diagonal entries are much too large for the independence hypothesis to be plausible.
Now "weight out" the entries of the table for which $i = j$. Now the deviance is much the same as its expected value, showing that ω is a good fit. From the table of "fitted values" we could find $\hat{\theta}$, and so on.
This model is what sociologists call a "mover-stayer" model; it is also an example of a "quasi-independence" model for a contingency table.

Bookprice-data: (this question now omitted, but see worksheet5)
Let p_{ij} be the price of the ith book in the jth country . Not all of the 50 such prices are available to us. We will fit models assuming that

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

with the usual assumption(to be checked later) that ϵ_{ij} are $NID(0, \sigma^2)$.
The question asks us to test $\beta = 0$, allowing for α. Now the deviances fitting
$E(p_{ij}) = \mu,\ \mu + \beta_j, \mu + \alpha_i + \beta_j, \mu + \alpha_i$
are respectively

$$684.56(df = 40), 532.29(df = 36), 134.84(df = 27), 227.54(df = 31).$$

So, to test $\omega : \beta = 0$, assuming Ω, we compute

$$[(dev(\omega) - dev(\Omega))/4]/[dev(\Omega)/27]$$

$= [92.70/4]/[134.84/27]$ and refer this to $F_{4,27}$.
Now the decrease in deviance in moving from the model $E(p_{ij}) = \mu + \beta_j$ to the model Ω is 397.5 (df=9), while 457.0(df=9) is the decrease in deviance in moving from the model $E(p_{ij}) = \mu$ to the model $E(p_{ij}) = \mu + \alpha_i$. The reason that these 2 quantities differ is because the parameters $(\alpha_i), (\beta_j)$ are not mutually orthogonal, since the "missing" values p_{ij} make the design unbalanced.
With the model

$$log(p_{ij}) = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

we find from the qplots that the assumption ϵ_{ij} dist'd as $NID(0, \sigma^2)$ is much more plausible than it was for the untransformed data. (Furthermore, models with MONEY as the response variable are far more likely to work in a multiplicative way than in an additive way: think of a percentage increase in your salary .)

10. These data on women's and men's tennis matches are taken from "An introduction to categorical data analysis " by Alan Agresti (1995) and their analysis is discussed in Agresti's Chapter 9.
i) The model being fitted is:
w_{ij} (the number of wins) is dist'd as independent $Bi(t_{ij}, p_{ij})$ for$1 \leq i < j \leq 5$ with $logit(p_{12}) = \mu + sel - graf, logit(p_{13}) = \mu + sel - sab$ etc
but the import of the -1 term in the glm statement is that we set $\mu = 0$.
Hence our model is $logit(p_{ij}) = \alpha_i - \alpha_j$ for $1 \leq i < j \leq 5$.
But note that with this model we could replace eg, α_5 by $\alpha_5 + 13$, and then replace α_4 by $\alpha_4 + 13$ and so on, without changing the formula for $logit(p_{ij})$.
So we impose a constraint, without loss of generality $\alpha_5 = 0$, to ensure **parameter identifiability** .
(If we include a term + sanc in our fitting, glm( ) will estimate sel,graf,saba,navr as given, but will obligingly tell us, in effect, that sanc is *aliased*, meaning that it cannot be estimated if the previous 4 parameters are already in the model.)
This is an example of the Bradley-Terry model for paired comparisons.
ii) Can we confidently say that Graf is better than Sanchez?
YES, because the model fits well (refer its deviance of 4.65 to χ^2_6) and $\hat{\alpha}_2/se(\hat{\alpha}_2) = 2.854$: refer this to $N(0, 1)$.
iii) Can we confidently say that Graf is better than Seles?
Now $\hat{\alpha}_2 - \hat{\alpha}_1 = 1.933 - 1.533$.
and the estimated variance of $(\hat{\alpha}_2 - \hat{\alpha}_1) = (.6687)^2$

(Use the parameter estimate correlation matrix)
Referring $(\hat{\alpha}_2 - \hat{\alpha}_1)/.6687 = 1.886$ to $N(0,1)$, we see that the difference is significant (but not spectacularly so).
iv) Our estimate of the probability that Sabatini beats Sanchez, in one match,is

$$e^{.7308}/(1+e^{.7308}) = 1/(1+e^{-.7308}) = .675$$

and note that using $.7308 \pm 1.96 \times .6764$
we could attach a confidence interval to this probability
ie $1/(1+exp(-.7308 \pm 1.96 \times .6764))$.
v) The data for the men's tennis doesn't fit the Bradley-Terry model so well.
Note that our underlying model contains the assumption
w_{ij} dist'd independently as $Bi(t_{ij}, p_{ij})$.
This may not be reasonable. For example, if Graf beats Sanchez in the first match, then this result may affect the probability that Graf beats Sanchez in their next match, and so on. Further, we should perhaps take account of the surface on which the match is played (grass, clay, etc). But we do not have the data to be able to take account of such factors in our modelling.

11. Take β_0 as the true value of β. The usual Taylor series expansion of the first derivative of $l_p(\beta)$ at $\hat{\beta}$ about β_0 shows that, to first order,

$$0 = U(\beta_0) - (\hat{\beta} - \beta_0)I(\beta_0)$$

which gives the required approximation for $(\hat{\beta} - \beta_0)$.
Now, as usual, $E(U(\beta_0)) = 0$, giving us the required expression for $E(\hat{\beta})$. Further, by noting that

$$var(\hat{\beta}) = var(\Sigma x_i Y_i)/(I(\beta_0))^2$$

and substituting $var(Y_i) = \phi\mu_i$, we obtain the required expression for $var(\hat{\beta})$.

....

For number 7 (lager data)

```
>lager <- read.table("lager", header=T)
>summary(lager)
>pairs(lager)
>attach(lager)
> first.lm <- lm(Price ~ og + percent + cal)
>summary(first.lm,cor=F)
Call: lm(formula = Price ~ og + percent + cal)
Residuals:
    Min     1Q  Median    3Q   Max
 -6.953 -3.217 -0.2989 1.827 16.53

Coefficients:
                 Value Std. Error    t value   Pr(>|t|)
(Intercept) -1430.4215  1487.7819    -0.9614     0.3427
         og     1.4393     1.4849     0.9693     0.3389
    percent    -2.4402     5.0846    -0.4799     0.6342
        cal    -0.3201     0.5135    -0.6233     0.5370

Residual standard error: 4.856 on 36 degrees of freedom
Multiple R-Squared: 0.4136
F-statistic: 8.465 on 3 and 36 degrees of freedom, the p-value is 0.0002186

>summary(first.lm,cor=T)
```

```
Call: lm(formula = Price ~ og + percent + cal)
Residuals:
    Min     1Q  Median    3Q   Max
 -6.953 -3.217 -0.2989 1.827 16.53

Coefficients:
                 Value Std. Error    t value   Pr(>|t|)
(Intercept) -1430.4215  1487.7819    -0.9614     0.3427
         og     1.4393     1.4849     0.9693     0.3389
    percent    -2.4402     5.0846    -0.4799     0.6342
        cal    -0.3201     0.5135    -0.6233     0.5370

Residual standard error: 4.856 on 36 degrees of freedom
Multiple R-Squared: 0.4136
F-statistic: 8.465 on 3 and 36 degrees of freedom, the p-value is 0.0002186

Correlation of Coefficients:
        (Intercept)      og percent
     og -1.0000
percent  0.4782     -0.4781
    cal  0.9217     -0.9218  0.1035

>next.lm <- lm(Price ~ og); summary(next.lm)
Call: lm(formula = Price ~ og)
Residuals:
    Min     1Q  Median    3Q   Max
 -7.617 -3.135 -0.3271 2.045 16.89

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept) -336.5331   71.2697    -4.7220    0.0000
         og    0.3475    0.0684     5.0802    0.0000

Residual standard error: 4.763 on 38 degrees of freedom
Multiple R-Squared: 0.4045
F-statistic: 25.81 on 1 and 38 degrees of freedom, the p-value is 1.033e-05
>Rat <- (rat>0)*1 ; Rat <- factor(rat)
>third.lm <- lm(Price ~ og + Rat); summary(third.lm)
Call: lm(formula = Price ~ og + Rat)
Residuals:
    Min     1Q Median    3Q   Max
 -7.609 -3.343  -0.22 1.934 17.05

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept) -330.0376   78.3167    -4.2141    0.0002
         og    0.3415    0.0748     4.5667    0.0001
        Rat   -0.3635    1.7006    -0.2138    0.8319

Residual standard error: 4.824 on 37 degrees of freedom
Multiple R-Squared: 0.4052
F-statistic: 12.6 on 2 and 37 degrees of freedom, the p-value is 6.696e-05

>last.lm <- lm(Price ~ Rat); summary(last.lm)
Call: lm(formula = Price ~ Rat)
Residuals:
```

```
    Min      1Q Median    3Q   Max
 -9.067 -4.405  -0.78 3.273 19.22

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  27.5667   1.5370    17.9359   0.0000
        Rat  -3.2867   1.9441    -1.6906   0.0991

Residual standard error: 5.953 on 38 degrees of freedom
Multiple R-Squared: 0.06995
F-statistic: 2.858 on 1 and 38 degrees of freedom, the p-value is 0.09911
>tapply(Price,Rat,mean)
        0     1
 27.56667 24.28
```

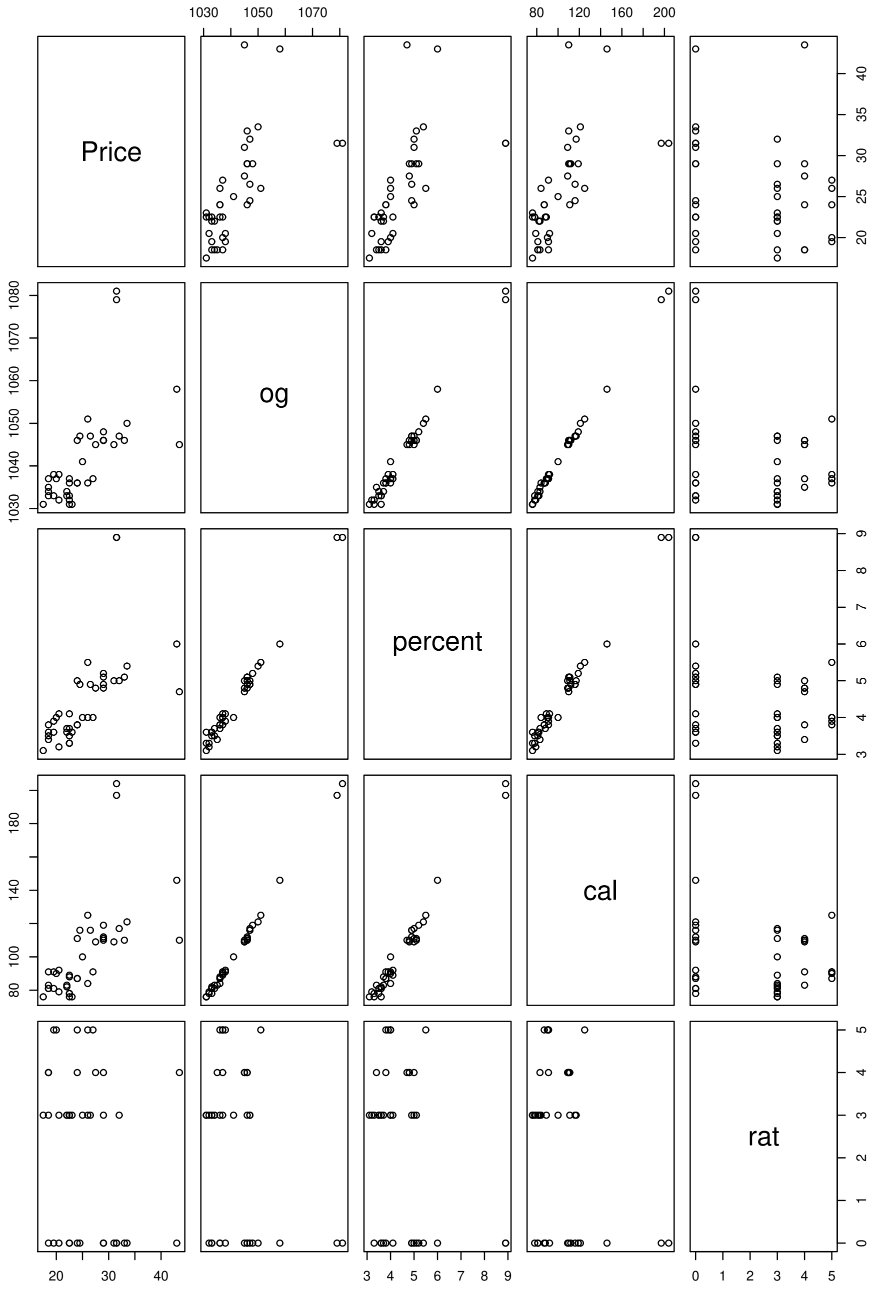

Figure 1: The pairs plot for the lager data