MATH3823 weekly exercises (with solutions)

Dr Stuart Barber

April 2022

These questions are intended to be worked on alongside reading the notes. To help keep track of where we are up to, they are separated by week they were set.

Week 2

1. *Verify that if* $q = 1/(1+e^{-x})$ *then* $x = \log(q/(1-q))$.

   Solving the first equation for x yields

$$\begin{aligned} & q = \frac{1}{1+e^{-x}} \\ \Leftrightarrow \quad & q(1+e^{-x}) = 1 \\ \Leftrightarrow \quad & e^{-x} = \frac{1-q}{q} \\ \Leftrightarrow \quad & x = \log\left(\frac{q}{1-q}\right), \end{aligned}$$

   as required.

2. *For the birthweight data example in section 1.2:*

   - *What obvious model have we not listed in the notes? Looking at the data plots in Figure 1.2, comment on whether you think this model is likely to be worth considering.*

     We did not attempt to fit the model `birthweight ~ sex`. This would be represented as two horizontal lines on the data plot, which does not look like a sensible explanation for the data so we do not pursue it further here.

   - *Test whether the* `age * sex` *interaction term is statistically significant for these data and comment on the resulting model.*

```
bwt = read.table("../data/birthwt.txt", header=T)
fit3 = lm(weight ~ age * sex, data = bwt)
coefficients(fit3)
```

```
## (Intercept)         age         sexM    age:sexM
## -2141.66667   130.40000   872.99425   -18.41724
```

```
anova(fit3)
```

```
## Analysis of Variance Table
##
## Response: weight
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## age        1 1013799 1013799 31.0779 1.862e-05 ***
## sex        1  157304  157304  4.8221   0.04006 *
## age:sex    1    6346    6346  0.1945   0.66389
## Residuals 20  652425   32621
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the R output, the p-value is 0.664 which indicates the interaction term is not statistically significant. We can worth this out step-by-step; the F-statistic is

$$F_{23} = \frac{(R_2 - R_3)/(r_2 - r_3)}{R_3/r_3} = \frac{6346/1}{658771/20} \approx 0.1926.$$

The corresponding F-distribution critical value and p-value can be obtained with

```
qf(0.95, 1, 20)
```

```
## [1] 4.351244
```

```
1-pf(0.1945, 1, 20)
```

```
## [1] 0.663928
```

Week 3

3. *Express each of the following distributions as an exponential family distribution, possibly with a scale parameter. Use the properties of exponential families to find the expectation and variance of each distribution.*

   - *The geometric distribution with probability mass function*

   $$f(y) = (1-p)^{y-1}p; \qquad y = 1, 2, 3 \ldots$$

   where $0 < p < 1$.

   The log pmf is

   $$\begin{aligned} \log f(y) &= (y-1)\log(1-p) + \log p, \quad y = 1, 2, 3, \ldots \\ &= y\log(1-p) + \log\{p/(1-p)\} \\ &= \theta y - b(\theta) \end{aligned}$$

where $\theta = \log(1-p),\ -\infty < \theta < 0,\ p = 1 - e^{\theta}$, and $b(\theta) = -\log(e^{-\theta} - 1) = \theta - \log(1 - e^{\theta})$. Also $1 - p = e^{\theta},\ p/(1-p) = e^{-\theta} - 1$. There is no scale parameter in this case. The mean and variance of Y are well-known and can also be found by differentiating $b(\theta)$,

$$E(Y) = b'(\theta) = 1 + \frac{e^{\theta}}{1 - e^{\theta}} = \frac{1}{1 - e^{\theta}} = 1/p,$$
$$\mathrm{var}(Y) = b''(\theta) = \frac{e^{\theta}}{(1 - e^{\theta})^2} = (1-p)/p^2.$$

- *The gamma distribution with probability density function*

$$f(y) = \Gamma(\alpha)^{-1}\lambda^{\alpha}y^{\alpha-1}e^{-\lambda y} \qquad y \geqslant 0,$$

where $\alpha, \lambda > 0$. (Hint: Treat $1/\alpha$ as a scale parameter and let θ be a suitable function of λ and α.)

This can be interpreted as a one-parameter exponential family with scale parameter $1/\alpha$. Write the log pdf (probability density function) as

$$\begin{aligned}\log f(y) &= \frac{-(\lambda/\alpha)y + \log\lambda}{1/\alpha} + (\alpha - 1)\log y - \log\Gamma(\alpha) \\ &= \frac{-(\lambda/\alpha)y + \log\lambda/\alpha}{1/\alpha} + (\alpha - 1)\log y - \log\Gamma(\alpha) + \alpha\log\alpha \\ &= \frac{\theta y - b(\theta)}{\varphi} + c(y, \varphi),\end{aligned}$$

for $y > 0$, where

$$\begin{aligned}\theta &= -\lambda/\alpha\ (< 0), \quad b(\theta) = -\log(-\theta) \\ \varphi &= 1/\alpha\ (> 0), \\ c(y,\varphi) &= (\alpha - 1)\log y - \log\Gamma(\alpha) + \alpha\log\alpha \\ &= \left(\frac{1-\varphi}{\varphi}\right)\log y - \log\Gamma(\varphi^{-1}) - \varphi^{-1}\log\varphi.\end{aligned}$$

(Note the extra term $\alpha\log\alpha$ which appears in $c(y,\varphi)$.) Then

$$E(Y) = b'(\theta) = \alpha/\lambda \quad \mathrm{var}(Y) = \varphi b''(\theta) = \alpha/\lambda^2.$$

(Take care differentiating $b(\theta)$). These moments for Y are well-known for the gamma distribution.

Week 4

4. *Find the canonical link functions for the geometric and gamma distributions.*

Let $\mu = \mathbb{E}(Y)$. Then:

- For the Geometric(p) distribution, we have $\mu = b'(\theta) = (1 - e^{\theta})^{-1}$. Hence

$$\begin{aligned} 1 - e^{\theta} &= 1/\mu \\ e^{\theta} &= 1 - 1/\mu \\ \theta &= \log(1 - 1/\mu) = g(\mu). \end{aligned}$$

- For the Gamma(α, λ) distribution, $\mu = b'(\theta) = -1/\theta$, so $g(\mu) = -1/\mu$.

Week 5

5. *Check that you can derive the formulae given for deviance of normal, binomial, and Poisson models just after equation (2.63) in the lecture notes.*

   - Under the normal model, with $\mu = \theta$, $b(\theta) = \theta^2/2$, $\phi = \sigma^2$, likelihood takes the form

$$l = \sum_{i=1}^{n} \{[\mu_i y_i - \mu_i^2/2]/\sigma^2 + c(y_i, \phi)\}$$

     let $\hat{l}$ denote the maximized likelihood under the generalized linear model of interest, with μ_i replaced by $\hat{\mu}_i$.

     Under the saturated model, μ_i is estimated by maximizing the ith term of the likelihood, yielding $\tilde{\mu} = y_i$. Substituting these values and noting that $y_i y_i - y_i^2/2 = y_i^2/2$ yields the maximized saturated log-likelihood

$$\tilde{l} = \sum_{i=1}^{n} \{[y_i^2/2]/\sigma^2 + c(y_i, \phi)\}$$

     Hence the deviance reduces to the residual sum of squares

$$D = 2\phi(\tilde{l} - \hat{l}) = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2.$$

   - Binomial data consist of the number of successes y_i out of m_i trials, $i = 1, \ldots, n$. The likelihood takes a form that is slightly modified from the general case,

$$\begin{aligned} l &= \sum_{i=1}^{n} \{\theta_i y_i - b_i(\theta_i) + c(y_i, m_i)\} \\ &= \sum_{i=1}^{n} \{y_i \log(p_i) + (m_i - y_i) \log(1 - p_i) + c(y_i, m_i)\}. \end{aligned}$$

     where $\theta_i = \text{logit}(p_i)$ and

$$p_i = \mu_i/m_i = E(Y_i)/m_i = b_i'(\theta_i)/m_i$$

is the probability of success in the ith set of trials, and $b_i(\theta_i) = m_i \log(1 + e^{\theta_i})$. Also,

$$\exp\{c(y_i, m_i)\} = \binom{m_i}{y_i}.$$

Under the saturated model p_i is estimated by $\tilde{p}_i = y_i/m_i$.
Hence the deviance takes the form

$$D = 2(\tilde{l} - \hat{l}) = 2\sum_{i=1}^{n} \left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i) \log\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right) \right\}.$$

In general $\hat{p}_i = \text{logit}^{-1}(\hat{\beta}^T x_i)$ will always lie strictly between 0 and 1. However, if $y_i = 0$ or m_i, then the saturated estimate $\tilde{p}_i$ will equal 0 or 1. Suppose that $y_i = 0$ for some i. Then the ith term of the saturated likelihood becomes $m_i + c(0, m_i)$. You should convince yourself that the formula for the deviance remains valid if we interpret $0 \log 0 = 0$. Similarly if $y_i = m_i$.

- In the Poisson case the likelihood becomes

$$\begin{aligned} l &= \sum_{i=1}^{n} \{\theta_i y_i - b(\theta_i) + c(y_i)\} \\ &= \sum_{i=1}^{n} \{y_i \log \lambda_i - \lambda_i + c(y_i)\}, \end{aligned}$$

where $\mu_i = \lambda_i = \log \theta_i$ and $b(\theta_i) = \exp(\theta_i)$. Under the saturated model $\tilde{\lambda}_i = y_i$. For most models of interest, it can be shown that $\sum \hat{\lambda}_i = \sum y_i$. Hence the deviance reduces to

$$D = 2(\tilde{l} - \hat{l}) = 2\sum_{i=1}^{n} \{y_i \log(y_i/\hat{\mu}_i)\}.$$

6. *(After reading section 2.6 of the notes) Go back to the birthweight example and consider how the deviance tests from section 2.6 relate to the tests we used to choose which model was the best fit to the data.*

In section 1.2, we analysed the birthweight data using F-tests based on the residual sum of squares R and residual degrees of freedom r' (note the change of notation — r was used for degrees of freedom in section 1.2, but later we have used r for the number of parameters). In section 2.6 we described tests using the deviance D and number of parameters r.

Since $D = R$ for a normal GLM and $r' = n - r$, where n is the number of observations, the F-statistic (1.8) comparing Model 0 and Model 1 can be written

$$F_{01} = \frac{(R_0 - R_1)/(r'_0 - r'_1)}{R_1/r'_1} = \frac{(D_0 - D_1)/(r_1 - r_0)}{D_1/(n - r_1)},$$

which we compared to an F-distribution on $r_1 - r_0$ and $n - r_1$ degrees of freedom. This is almost equivalent to the test statistic (2.66):

$$\frac{(D_1 - D_2)/(r_2 - r_1)}{D_3/(n - r_3)},$$

which we compared to an $F_{r_2-r_1,n-r_3}$ distribution.

The only differences are (i) the arbitrary indexing of models (0,1) or (1,2) and (ii) that in the second form we use a "large" Model 3 to estimate σ^2 by $\hat{\phi} = D_3/(n - r_3)$. In practice the differences will usually be small and the two approaches to variable selection are equivalent.

Week 6

7. *Complete the exercises in the R script from lecture 4:*

   - *Check the two different ways of fitting a Poisson GLM to the toy data give the same result;*

     A wider question is what it means to say two models are "the same"? If you have the same type of model (Poisson GLM) using the same covariates, then it is enough to check whether the coefficients are the same. Assuming that the fitted models are `y.glm` and `y.glm2`, then you can check this with any of the following:

```
cbind(coef(y.glm), coef(y.glm2)) # detailed comparison
```

```
##                   [,1]       [,2]
## (Intercept)  1.1408451  1.1408451
## a2          -0.3053816 -0.3053816
## a3           0.1466035  0.1466035
## a4           0.4567584  0.4567584
## b2           0.9162907  0.9162907
## b3           0.9444616  0.9444616
```

```
coef(y.glm) - coef(y.glm2) # direct comparison
```

```
## (Intercept)          a2          a3          a4          b2          b3
##           0           0           0           0           0           0
```

```
range(coef(y.glm) - coef(y.glm2)) # useful for comparing larger objects
```

```
## [1] 0 0
```

   - *Check whether a smaller model is adequate for the carbohydrate uptake in diabetes data set;*

     The current model is `carb ~ weight + protein`, with both `weight` and `protein` being continuous variables so they have one degree of freedom each. The two simpler models

to consider are `carb ~ weight` and `carb ~ protein`. The second one has the lower deviance, so we test H_0: `carb ~ protein` is sufficient against H_A: `carb ~ weight + protein` is preferred with the following *R* code (assuming the script from lecture 4 has already been run).

```
glm.p = glm(carb ~ protein)
dev.p = deviance(glm.p); df.p = df.residual(glm.p)
f.p.wp = ((dev.p-dev.wp)/(df.p-df.wp)) / (dev.3way/df.3way)
pf(f.p.wp, (df.p-df.wp), df.3way, lower.tail=F)
```

```
## [1] 0.01671717
```

Since the p-value is less than 0.05, we reject H_0 and conclude that we can't simplify the model further than `carb ~ weight + protein`.

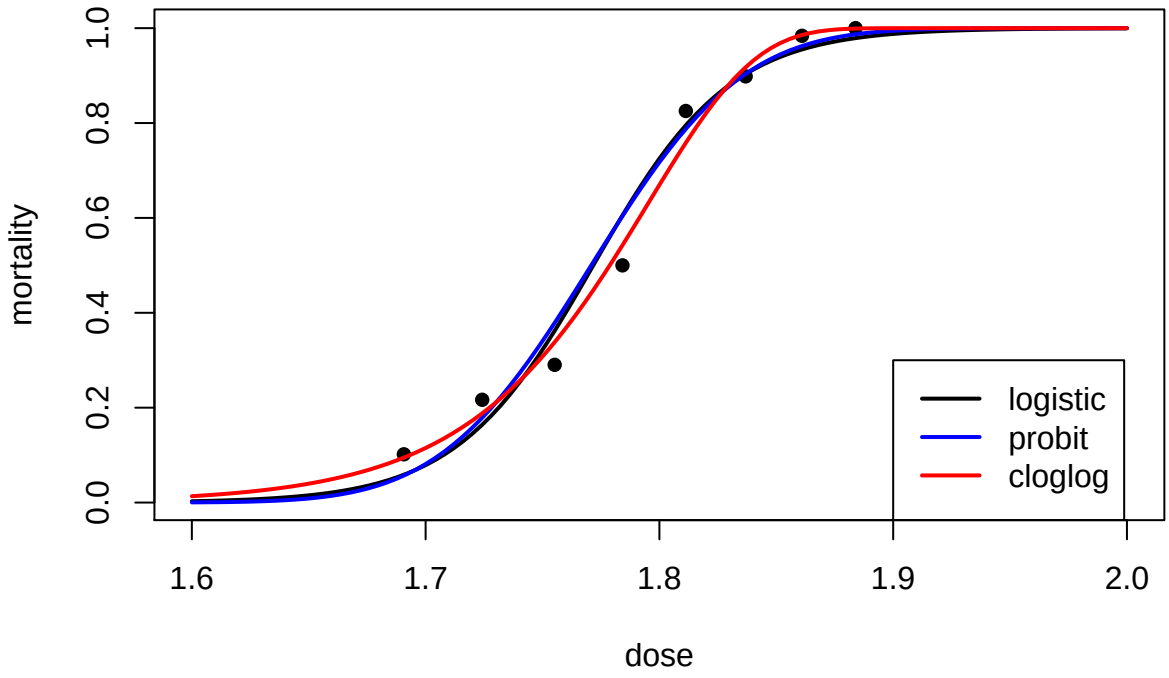
- *Fit alternative link functions to the beetle data and plot the results.*

  Assuming you've already run the script to plot the logistic model, these commands will redraw the plot with probit and complementary log-log curves added.

```
## Fit probit and cll models, evaluate predicted probabilities
beetle.glm.p = glm(ym ~ dose, family=binomial(link="probit"))
beetle.glm.c = glm(ym ~ dose, family=binomial(link="cloglog"))
pred.p = predict(beetle.glm.p, newdata = newdata, type="response")
pred.c = predict(beetle.glm.c, newdata = newdata, type="response")

## Re-draw plot with added lines
plot(mortality ~ dose, pch=16, xlim=range(newdata$dose),
     ylim=range(pred))
lines(pred ~ newdata$dose, lwd=2)
lines(pred.p~newdata$dose, col="blue", lwd=2)
lines(pred.c~newdata$dose, col="red", lwd=2)

## Label plot
legend(x=1.9, y=0.3, lwd=2, lty=1,
       col=c("black", "blue", "red", "white"),
       legend=c("logistic", "probit", "cloglog", ""))
```



8. *Use the fitted logistic regression model to predict what dose of gaseous carbon disulphide would kill 90% of beetles.*

   The model connects probability of death p to dose x via

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \eta(x) = \beta_0 + \beta_1 x.$$

   Hence the required dose for any p is $x(p) = (\text{logit} - \beta_0)/\beta_1$. Substituting the given $p = 0.9$ and fitted $\hat{\beta}_0 = -60.717$ and $\hat{\beta}_1 = 34.270$, we obtain

$$x(0.9) = (\log(9) + 60.717)/34.270 = 1.83584.$$

Week 8

9. *Analyse the colon cancer data available in Minerva. Specifically, use a logistic regression model to investigate whether the genetic mutation and/or the environmental exposure affect the incidence of colon cancer.*

   R hints:

   - *encode 'agecat' as a factor, as shown in the 'data' folder on Minerva.*

- *Note that the response variable for a binomial glm in R can be a vector (not a matrix) in the special case of $m = 1$, so a formula of the form 'case ~ ....' will work for this data set.*

First, we load and prepare the data.

```
## Load data and collapse exposure to 0/1
gs = read.table("genestudy.txt", header=T)
gs$exp = as.numeric(gs$exp>0)
table(read.table("genestudy.txt", header=T)$exp, gs$exp)
```

```
##
##       0   1
##   0 280   0
##   1   0 329
##   2   0 162
```

```
## Create alternative version with factor variables
gsf = gs
gsf$agecat = as.factor(gsf$agecat)
gsf$gene = as.factor(gsf$gene)
gsf$exp = as.factor(gsf$exp)
```

Next, fit a range of models and note their deviances and degrees of freedom.

```
attach(gsf)
fit.age = glm(case ~ agecat + gene + exp, family=binomial)
fit.ge = glm(case ~ gene + exp, family=binomial)
fit.ae = glm(case ~ agecat + exp, family=binomial)
fit.ag = glm(case ~ agecat + gene, family=binomial)
fit.a = glm(case ~ agecat, family=binomial)
fit.g = glm(case ~ gene, family=binomial)
fit.e = glm(case ~ exp, family=binomial)
fit.0 = glm(case ~ 1, family=binomial)

dev = c(deviance(fit.age), deviance(fit.ge), deviance(fit.ae), deviance(fit.ag), devianc

deg = c(df.residual(fit.age), df.residual(fit.ge), df.residual(fit.ae), df.residual(fit.

tmp = cbind(dev, deg)
rownames(tmp) = c("age", "ge", "ae", "ag", "a", "g", "e", "null")
tmp
```

```
##           dev deg
## age  725.7856 764
## ge   739.9060 768
## ae   727.5703 765
```

```
## ag   726.0892 765
## a    727.8718 766
## g    740.2998 769
## e    742.0783 769
## null 742.4823 770
```

```
qchisq(0.05, 1:5, lower.tail=F)
```

```
## [1]  3.841459  5.991465  7.814728  9.487729 11.070498
```

Whether we start with the null model and work up or the full main effects model and work down, variable selection by deviance tests results in choosing the model `case ~ age`. It appears that the incidence of colon cancer is affected by age category, but not by carrying the genetic mutation or chemical exposure. Inspecting this model gives us the following results.

```
summary(fit.a)
```

```
##
## Call:
## glm(formula = case ~ agecat, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7747  -0.7350  -0.5533  -0.5230   2.0284
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.72432    0.22634  -7.618 2.57e-14 ***
## agecat1     -0.19606    0.34449  -0.569   0.5693
## agecat2     -0.07499    0.32281  -0.232   0.8163
## agecat3      0.67450    0.28874   2.336   0.0195 *
## agecat4      0.55339    0.28980   1.910   0.0562 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 742.48  on 770  degrees of freedom
## Residual deviance: 727.87  on 766  degrees of freedom
## AIC: 737.87
##
## Number of Fisher Scoring iterations: 4
```

We can estimate the probability of developing colon cancer as follows.

```
summary(fit.a)
```

```
##
```

```
## Call:
## glm(formula = case ~ agecat, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7747  -0.7350  -0.5533  -0.5230   2.0284
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.72432    0.22634  -7.618 2.57e-14 ***
## agecat1     -0.19606    0.34449  -0.569   0.5693
## agecat2     -0.07499    0.32281  -0.232   0.8163
## agecat3      0.67450    0.28874   2.336   0.0195 *
## agecat4      0.55339    0.28980   1.910   0.0562 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 742.48  on 770  degrees of freedom
## Residual deviance: 727.87  on 766  degrees of freedom
## AIC: 737.87
##
## Number of Fisher Scoring iterations: 4
```

```
beta = coef(fit.a)
eta = beta[1] + c(0, beta[-1]) # lin pred values for each group
pred = exp(eta) / (1 + exp(eta))
names(pred)[1] = "agecat0"
pred
```

```
##   agecat0   agecat1   agecat2   agecat3   agecat4
## 0.1513158 0.1278195 0.1419355 0.2592593 0.2366864
```

Note that these are just the same as the observed proportion of cases in each age group, as shown in table 1.

Table 1: Probabilities of developing colon cancer in each age category.

	agecat0	agecat1	agecat2	agecat3	agecat4
modelled	0.151	0.128	0.142	0.259	0.237
observed	0.151	0.128	0.142	0.259	0.237

Week 11

10. *Overdispersion is an occasional problem when fitting generalised linear models with a known scale parameter in situations where there is unexplained variation. In this exercise we illustrate the problem in Poisson regression. The starting point is the observation that if* $Y \sim P(\lambda)$*, then* $E(Y) = \lambda$ *and* $\mathrm{var}(Y) = \lambda$*.*

    a. *Consider joint random variables* (X, Y) *where* X *takes two possible values with equal probabilities,* $\Pr(X = 1) = \Pr(X = 2) = 1/2$*. Suppose the conditional distribution of* Y *given* X *is Poisson,*

    $$Y|X = 1 \sim P(\lambda_1), \quad Y|X = 2 \sim P(\lambda_2), \qquad (*)$$

    where $\lambda_1 < \lambda_2$*. Let* $\lambda = (\lambda_1 + \lambda_2)/2$ *denote the average value. Thus the marginal distribution of* Y *is a mixture of two Poisson distributions. Show that*

    $$E(Y) = \lambda, \quad \mathrm{var}(Y) = \lambda + (\lambda_1 - \lambda_2)^2/4,$$

    that is, although the mean of Y *is the same under the mixture (or conditional Poisson) model as under the Poisson model, the variance is larger.*

    If $Y \sim P(\lambda)$, then $E(Y) = \lambda,\ \mathrm{var}(Y) = \lambda,\ E(Y^2) = \lambda + \lambda^2$. Hence for the mixture model

    $$E(Y) = E(Y|X = 1)P(X = 1) + E(Y|X = 2)P(X = 2) = \frac{1}{2}(\lambda_1 + \lambda_2)$$
    $$E(Y^2) = E(Y^2|X = 1)P(X = 1) + E(Y^2|X = 2)P(X = 2) = \frac{1}{2}(\lambda_1 + \lambda_1^2 + \lambda_2 + \lambda_2^2),$$

    so that

    $$\begin{aligned}\mathrm{var}(Y) &= \frac{1}{2}(\lambda_1 + \lambda_1^2 + \lambda_2 + \lambda_2^2) - \left\{\frac{1}{2}(\lambda_1 + \lambda_2)\right\}^2 \\ &= \lambda + (\lambda_1 - \lambda_2)^2/4.\end{aligned}$$

    b. *This phenomenon might be observed in data as follows. Let* $n = 60$ *and let an explanatory variable* x_i *take the value* $x_i = 1$ *for* $i = 1, \ldots, 30$ *and* $x_i = 2$ *for* $i = 31, \ldots, 60$*. Suppose that the observations* $y_i|x_i$ *come from the above conditional Poisson model* $(*)$*.*

    Consider fitting the following two models in R with Poisson errors and a log link function:

    $$\text{(i)}\ \ y \sim 1, \qquad \text{(ii)}\ \ y \sim x.$$

    Since model (ii) is the correct model, it should yield a good fit to the data. But if the experimenter does not know about the variable x*, it will only be feasible to fit model (i). Let* $\overline{Y}$ *and* S^2 *denote the sample mean and variance of the* $\{Y_i\},\ i = 1, \ldots, 60$*. Show that*

    $$E(\overline{Y}) = \lambda, \quad E(S^2) = \lambda + \frac{60}{59}(\lambda_1 - \lambda_2)^2/4.$$

Hence show that the χ^2 goodness of fit statistic for model (i) will indicate a poorly fitting model if λ_1 and λ_2 are far apart.

Writing $\overline{Y} = \frac{1}{2}(\overline{Y}_1 + \overline{Y}_2)$, where $\overline{Y}_j$ has mean λ_j and variance $\lambda_j^2/30$, $j = 1, 2$, we see that

$$E\{(\overline{Y})^2\} = \frac{1}{4}E\left\{(\overline{Y}_1)^2 + 2\overline{Y}_1\overline{Y}_2 + (\overline{Y}_2)^2\right\} = \frac{1}{4}\{\lambda_1^2 + \lambda_1/30 + 2\lambda_1\lambda_2 + \lambda_2^2 + \lambda_2/30\}.$$

Also

$$E\{\sum_{i=1}^{60} Y_i^2\} = E\{\sum_{i=1}^{30} Y_i^2\} + E\{\sum_{i=31}^{60} Y_i^2\} = 30(\lambda_1^2 + \lambda_1 + \lambda_2^2 + \lambda_2).$$

Hence

$$\begin{aligned} E(S^2) &= \frac{1}{59}E\left\{\sum_{i=1}^{60} Y_i^2 - 60(\overline{Y})^2\right\} \\ &= \frac{1}{2}(\lambda_1 + \lambda_2) + \frac{60}{4 \cdot 59}(\lambda_1 - \lambda_2)^2, \end{aligned}$$

which is larger than λ when $\lambda_1 \neq \lambda_2$.

The χ^2 goodness of fit statistic is just

$$X^2 = \sum(o - e)^2/e = (n - 1)S^2/\overline{Y},$$

in terms of observed values $o = o_i = Y_i$ and expected values $e = e_i = \overline{Y}$. Here $n = 60$. Under the null Poisson model, the numerator and denominator have the same expectation and for large n, $X^2 \sim \chi^2_{n-1}$ approximately. Under the mixture model the numerator has a larger expectation and X^2 will be larger in distribution than χ^2_{n-1}.

c. *This example is very simple, but overdispersion can occur much more widely. Why is overdispersion not a problem for generalised linear models in which the response distribution includes a scale parameter?*

When a scale parameter ϕ is present, it can be used to represent the true scale variability in the model, plus the effects of overdispersion.

11. *(Based on Dobson & Barnett, pp 144–145). Suppose there are 2 groups of people: the first group is exposed to some pollutant and the second group is not. In a prospective study, each group is followed for several years and categorized according to the presence or absence of some disease. Let π_i denote the probability that a person in group i contracts the disease, $i = 1, 2$. The following 2×2 table summarizes the different possibilities.*

	Diseased	Not diseased
Exposed	π_1	$1 - \pi_1$
Not exposed	π_2	$1 - \pi_2$

Note that the sum of each row is 1. For each $i = 1, 2$, the odds *of contracting the disease is defined by*

$$O_i = \pi_i/(1 - \pi_i),$$

and a comparison between these two probabilities is given by the odds ratio

$$\phi = \frac{O_1}{O_2} = \frac{\pi_1(1 - \pi_2)}{\pi_2(1 - \pi_1)}.$$

a. *Show that $\phi = 1$ if and only if there is no difference between the control and exposed groups. What does it mean if $\phi > 1$?*

We will have $\phi = 1$ if and only if $\{\pi_1/(1-\pi_1)\}/\{\pi_2/(1-\pi_2)\} = 1$ iff $\pi_1/(1-\pi_1) = \pi_2/(1-\pi_2)$. Since the "odds" map $\pi \to \pi/(1-\pi)$ is monotone from $(0, 1)$ to $(0, \infty)$ (check its derivative is everywhere positive), this last inequality holds iff $\pi_1 > \pi_2$.

b. *Consider now m 2×2 tables of this form, $j = 1, \ldots, m$, with probabilities π_{ij} represented by a logistic model*

$$\text{logit}(\pi_{ij}) = \alpha_i + \beta_i x_j, \ i = 1, 2, \ j = 1, \ldots, m,$$

where x_j is some specified quantitative explanatory variable. Interpret the parameters α_i and β_j, and give their effect on the log odds ratio $\log \phi_j$, say, for each table. Show that $\log \phi_j$ is constant across the m tables if $\beta_1 = \beta_2$.

The log odds ratio $\log\ \phi_j$ can be written as

$$\begin{aligned} \log\ \phi_j &= \text{logit}\, \pi_{1j} - \text{logit}\, \pi_{2j} \\ &= \alpha_1 - \alpha_2 + (\beta_1 - \beta_2)x_j, \quad j = 1, ..., m, \end{aligned}$$

which is constant in j iff $\beta_1 - \beta_2 = 0$.

The parameters α_i and β_i describe the intercept and slope of a logistic regression of π_{ij} on x_j. Thus the log odds ratio also varies linearly with x_j, with slope $\alpha_1 - \alpha_2$ and intercept $\beta_1 - \beta_2$.

c. *Give a practical example where such a model might be appropriate.*

Here is an example where this model might be plausible. Suppose a group of workers has been exposed to a dangerous pollutant, and this group has been matched to an unexposed control group. Further, suppose the smoking habits of each worker are recorded and given a score $x = 0, 1, 2$ (for none, medium, high). The workers are followed up for 40 years and the numbers dying of lung cancer are recorded. Let π_{ij} denote the probability of dying from lung cancer for workers in group i with smoking level j, $i = 1, 2,\ j = 0, 1, 2$. Then, with risk measured in terms of logit π_{ij},

- α_1 and α_2 represent the risk to nonsmokers in the exposed and unexposed groups. We might expect $\alpha_1 > \alpha_2$.
- β_1 and β_2 represent the increased risk as x increases by 1 for each group. We expect $\beta_1, \beta_2 > 0$.

If the risks of smoking act "independently" from those of exposure, we expect $\beta_1 = \beta_2$. But if there is a biological interaction between pollution and smoking, we might expect $\beta_1 > \beta_2$.

d. *How would you express this model in the R computer language?*

Let `count` be a $2m \times 2$ matrix containing the numbers of diseased and non-diseased people for each combination of i and j. Let `status` be a 2-level factor giving the group information i and let `x` be a quantitative variable storing the x_j information. This model can be fitted by the command `glm(count ~ status*x, binomial)`.

12. *Consider a $2 \times m$ contingency table with entries y_{ij}, $i = 1, 2,\ j = 1, \ldots, m$. The rows label STATUS (= 1,2 for alive/dead) and the columns label AGE groups (= 1,. . . ,m). Consider a Poisson model $P(\lambda_{ij})$ with*

$$\log(\lambda_{ij}) = \delta + \alpha_i + \beta_j + \gamma_i j$$

and discuss suitable aliasing conditions on the parameters (note that j is treated as a quantitative variable in the last term). Show that this model, conditioned on the column totals y_{+j}, reduces to a product binomial model $B(y_{+j}, \pi_j)$, $j = 1, \ldots, m$, and find the form of π_j. Which parameters are estimable under the binomial model?

Let `status` ($i = 1, 2$) and `age` ($j = 1, ..., m$) be factors, and let `y` store the entries in the contingency table. Also, let `agen` be a quantitative or numeric variable containing the same numbers as `age`. This model can be fitted in R by the command `glm(y ~ status + age + status:agen, poisson)` with parameters α_i, β_j and γ_i, $i = 1, 2;\ j = 1, \ldots, m$. The aliasing constraints are $\alpha_1 = 0$, $\beta_1 = 0$ and $\gamma_1 = 0$ (since the effects of $\gamma_1 j$ can be absorbed into μ and the β_j's).

Conditioning on y_{+j}, we see that the probability function of y_{1j} is given by

$$\begin{aligned} P(y_{1j} = k | y_{+j} = n) &= P(y_{1j} = k \text{ and } y_{2j} = n - k | y_{+j} = n) \\ &= \frac{e^{-\lambda_{ij}}(\lambda_{ij})^k e^{-\lambda_{2j}}(\lambda_{2j})^{n-k}/(k!(n-k)!)}{e^{-\lambda_{+j}}(\lambda_{+j})^n/n!} \\ &= \binom{n}{k} \pi_j^k (1-\pi_j)^{n-k}, \quad 0 \le k \le n \end{aligned}$$

where $\pi_j = \lambda_{1j}/\lambda_{+j}$. That is $y_{1j} \sim B(y_{+j}, \pi_j)$, where

$$\begin{aligned} \operatorname{logit} \pi_j = \log \lambda_{1j} - \log \lambda_{2j} &= \alpha_1 - \alpha_2 + (\gamma_1 - \gamma_2)j \\ &= -\alpha_2 - \gamma_2 j. \end{aligned}$$

Further, the $\{y_{ij},\ j = 1, \ldots, m \text{ given } y_{+j},\ j = 1, \ldots, m\}$ are independent. Thus the binomial data follow a logistic regression with intercept $-\alpha_2$ and slope $-\gamma_2$, which are the identifiable parameters.

13. *(From Dobson & Barnett, p 163) This question should be done in a computer package such as R. You should think carefully about which variables, if any, to condition on in your analysis: HOME = 1,2,3, CONTACT = 1,2, or SATISFACTION = 1,2,3.*

    The data relate to an investigation into satisfaction with housing conditions in Copenhagen. Residents of selected areas living in rented houses built between 1960 and 1968 were questioned about their satisfaction and their degree of contact with other residents. The data were tabulated by type of housing. Investigate the associations between satisfaction, contact with other residents and type of housing.

    Low Contact:

Satisfaction:	Low	Medium	High
Tower blocks	65	54	100
Apartments	130	76	111
Houses	67	48	62

    High Contact:

Satisfaction:	Low	Medium	High
Tower blocks	34	47	100
Apartments	141	116	191
Houses	130	105	104

    a. *Produce appropriate tables of percentages to gain initial insights into the data; for example, percentages in each contact level by type of housing and level of satisfaction, or percentages in each level of satisfaction by contact and type of housing.*

b. *Using e.g. R, fit various log-linear models to investigate interactions between the variables.*

c. *For some model that fits (at least moderately) well, calculate the Pearson residuals and use them to find where the largest discrepancies are between the observed and expected values.*

The three tables at the end give percentages adding to 100 over housing, satisfaction and contact, respectively. The R output shows the result of fitting a model with all 3 first-order interactions but no second-order interaction. The deviance $6.89 \sim \chi^2_4$ is not significant, so there is no need to consider the saturated model. The residuals all lie between ± 2, again indicating the model is an adequate fit. There are strongly significant coefficients within each first-order interaction; hence, further simplification has not been attempted. For this dataset it does not seem reasonable to condition on any of the marginal totals so the counts are assumed to arise from a Poisson model, not a product-multinomial model.

Since there is no second-order interaction, the log odds ratios for any two variables are constant given the third. Here are some interpretations based on the tables of percentages and the coefficients in the R output.

i. For each level of contact, Tables I and II show that the ratio of high to low satisfaction is higher for tower block residents than for apartment dwellers, which is in turn higher than for house residents. This interpretation is confirmed by the coefficients `sat3.housing2 = -.64` and `sat3.housing3 = -.95`.

ii. For each level of satisfaction, Tables I and III show that contact with neighbours increases as one moves from housing category tower block to apartment to house. This interpretation is confirmed by the coefficients `housing2.contact2 = .57` and `housing3.contact2 = .89`.

iii. For each level of housing, Tables II and III show that high satisfaction goes with high contact. This interpretation is confirmed in particular by the coefficient `sat3.contact2 = .33`.

For me the overall conclusions are partly expected and partly unexpected. I am surprised that tower blocks have a higher satisfaction than houses, and that house residents have greater contact than tower block residents. But it is natural that higher satisfaction is associated with higher contact.

I Percentages adding to 100% across housing

Low Contact:

Satisfaction:	Low	Medium	High
Tower blocks	25	30	37
Apartments	50	43	41
Houses	26	27	23
Total	100	100	100

High Contact:

Satisfaction:	Low	Medium	High
Tower blocks	11	18	25
Apartments	46	43	48
Houses	43	39	26
Total	100	100	100

II Percentages adding to 100% across housing

Low Contact:

Satisfaction:	Low	Medium	High	Total
Tower blocks	30	25	46	100
Apartments	41	24	35	100
Houses	38	27	35	100

High Contact:

Satisfaction:	Low	Medium	High	Total
Tower blocks	19	26	55	100
Apartments	31	26	43	100
Houses	38	31	31	100

III Percentages adding to 100% across contact

Low Satisfaction:

Contact:	Low	High	Total
Tower blocks	66	34	100
Apartments	48	52	100
Houses	34	66	100

Medium Satisfaction:

Contact:	Low	High	Total
Tower blocks	53	47	100
Apartments	40	60	100
Houses	31	69	100

High Satisfaction:

Contact:	Low	High	Total
Tower blocks	50	50	100
Apartments	37	63	100
Houses	37	63	100

```
## Enter data and construct data frame
count = c(65,54,100,34,47,100,130,76,111,141,116,191,67,48,62,130,105,104)
sat = rep(1:3,6)
housing = rep(1:3,rep(6,3))
contact = rep(rep(1:2,rep(3,2)),3)
sat = as.factor(sat)
housing = as.factor(housing)
contact = as.factor(contact)
data = cbind(count, sat, housing, contact)
colnames(data) = c("count", "sat", "housing", "contact")
data
```

```
##       count sat housing contact
##  [1,]    65   1       1       1
##  [2,]    54   2       1       1
##  [3,]   100   3       1       1
##  [4,]    34   1       1       2
##  [5,]    47   2       1       2
##  [6,]   100   3       1       2
```

```
##  [7,]   130   1       2       1
##  [8,]    76   2       2       1
##  [9,]   111   3       2       1
## [10,]   141   1       2       2
## [11,]   116   2       2       2
## [12,]   191   3       2       2
## [13,]    67   1       3       1
## [14,]    48   2       3       1
## [15,]    62   3       3       1
## [16,]   130   1       3       2
## [17,]   105   2       3       2
## [18,]   104   3       3       2
```

```
## Construct tables of percentages
m = matrix(count,ncol=6,byrow=T)
m2 = t(rbind(m[,1:3],m[,4:6]))
m2 = round(m2%*%diag(1/(as.vector(rep(1,3)%*%m2)))*100)
m2 = t(m2); m2 = cbind(m2[1:3,],m2[4:6,])
m2# data scaled to sum to 100 over satisfaction
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]   30   25   46   19   26   55
## [2,]   41   24   35   31   26   43
## [3,]   38   27   35   38   31   31
```

```
m3 = rbind(m[,c(1,4)],m[,c(2,5)],m[,c(3,6)])
m3 = t(m3); m3 = round(m3%*%diag(1/(as.vector(rep(1,2)%*%m3)))*100)
m3 = t(m3); m3 = cbind(m3[1:3,1],m3[4:6,1],m3[7:9,1],m3[1:3,2],m3[4:6,2],m3[7:9,2])
m3 # data scaled to sum to 100 over contact
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]   66   53   50   34   47   50
## [2,]   48   40   37   52   60   63
## [3,]   34   31   37   66   69   63
```

```
## Modelling
glm1 = glm(count ~ sat + housing + contact + sat:housing + contact:housing +
               sat:contact, poisson)
summary(glm1)
```

```
##
## Call:
## glm(formula = count ~ sat + housing + contact + sat:housing +
##     contact:housing + sat:contact, family = poisson)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
```

```
##  0.63752   0.01457  -0.50277  -0.81956  -0.01560   0.52021   0.37497   0.08952
##        9        10        11        12        13        14        15        16
## -0.46821  -0.35261  -0.07205   0.36546  -1.08007  -0.12692   1.36249   0.82987
##       17        18
##  0.08659  -0.96174
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         4.0943     0.1127  36.338  < 2e-16 ***
## sat2               -0.1073     0.1524  -0.704 0.481589
## sat3                0.5608     0.1329   4.219 2.46e-05 ***
## housing2            0.7402     0.1302   5.687 1.30e-08 ***
## housing3            0.2395     0.1417   1.690 0.090995 .
## contact2           -0.4306     0.1293  -3.331 0.000867 ***
## sat2:housing2      -0.4068     0.1713  -2.375 0.017570 *
## sat3:housing2      -0.6416     0.1501  -4.275 1.91e-05 ***
## sat2:housing3      -0.3371     0.1804  -1.869 0.061627 .
## sat3:housing3      -0.9456     0.1645  -5.749 8.98e-09 ***
## housing2:contact2   0.5744     0.1256   4.575 4.76e-06 ***
## housing3:contact2   0.8906     0.1387   6.419 1.37e-10 ***
## sat2:contact2       0.2960     0.1301   2.275 0.022909 *
## sat3:contact2       0.3282     0.1182   2.777 0.005483 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 294.477  on 17  degrees of freedom
## Residual deviance:   6.893  on  4  degrees of freedom
## AIC: 148
##
## Number of Fisher Scoring iterations: 4
```

```
residuals(glm1,type="pearson")
```

```
##           1           2           3           4           5           6
##  0.64620407  0.01457774 -0.49864032 -0.80142840 -0.01559242  0.52481845
##           7           8           9          10          11          12
##  0.37705287  0.08967078 -0.46480966 -0.35088696 -0.07196879  0.36708512
##          13          14          15          16          17          18
## -1.05756656 -0.12654027  1.40479462  0.84025074  0.08670781 -0.94719781
```