Count data, version 2

1 A data set

From `?housing` after `library(MASS)`: The housing data frame has 72 rows and 5 variables, cross-classifying 1681 individuals by:

Sat. Satisfaction of householders with their present housing circumstances, (High, Medium or Low, ordered factor).

Infl. Perceived degree of influence householders have on the management of the property (High, Medium, Low).

Type. Type of rental accommodation, (Tower, Atrium, Apartment, Terrace).

Cont. Contact residents are afforded with other residents, (Low, High).

Freq. Frequencies: the numbers of residents in each class.

```
head(housing)

##      Sat   Infl  Type Cont Freq
## 1    Low    Low Tower  Low   21
```

```
## 2 Medium    Low Tower  Low   21
## 3   High    Low Tower  Low   28
## 4    Low Medium Tower  Low   34
## 5 Medium Medium Tower  Low   22
## 6   High Medium Tower  Low   36
```

Regard `Sat` as the response and `Infl`, `Type`, and `Cont` as predictors. A simple model: each individual has a fixed probability of ending up in each response category,

$$\begin{aligned} p_{i,t,,c,s} &= \mathbb{P}\{Sat = s, Infl = i, Type = t, Cont = c\} \\ &= \mathbb{P}\{Infl = i, Type = t, Cont = c\} \times \\ &\qquad \mathbb{P}\{Sat = s \mid Infl = i, Type = t, Cont = c\} \\ &= \gamma_{itc} \times p_{i,t,c}(s) \end{aligned}$$

and individuals behave independently. For interpretation we are mostly interested in the conditional probabilities $p_{i,t,c}(s)$.

For `glm()`, the quantities $\log(p_{icts})$ are modelled as linear functions of the predictors, which are estimated by maximum likelihood. That is, the estimators are chosen to maximize the likelihood function

$$\mathcal{L}_{1681} = \prod_{\alpha=1}^{1681} p_{i_\alpha,t_\alpha,c_\alpha}(s_\alpha)\gamma_{i_\alpha,t_\alpha,c_\alpha},$$

where $(s_\alpha, i_\alpha, t_\alpha, c_\alpha)$ are the observed levels of the factors for individual α. Of course each $p_{i_\alpha,t_\alpha,c_\alpha}(s_\alpha)\gamma_{i_\alpha,t_\alpha,c_\alpha}$ needs to be rewritten as functions of the unknown parameters.

The housing data set does not give the individual responses. Luckily the likelihood only depends on aggregated counts. If

$$N_{itcs} = Freq[Sat = s, Infl = i, Type = t, Cont = c]$$

denotes the number of individuals for which $i_\alpha = i, t_\alpha = t, c_\alpha = c, s_\alpha = s$ then

$$\log \mathcal{L}_{1681} = \sum\nolimits_{s,i,t,c} N_{itcs} \left(\log p_{itc}(s) + \log \gamma_{itc}\right).$$

With count data of this form it is common to model the sample size $N = \sum_{i,t,c,s} N_{itcs}$ as random, with a Poisson(λ) distribution. That is,

$$\mathbb{P}\{N = n\} = e^{-\lambda}\frac{\lambda^n}{n!} \qquad \text{for } n = 0, 1, \ldots.$$

(For the housing data the observed N equals 1681.) Under this model the N_{itcs}'s become independent Poisson random variables, with expected values $\lambda p_{itcs}(s)$. Fortunately, the log-likelihood when $N = n$ is only slightly different from $\log \mathcal{L}_n$:

$$\text{log-likelihood} = -\lambda + n \log \lambda - \log(n!) + \mathcal{L}_n.$$

The $\widehat{p}_{itcs}$'s under the Poisson model are the same as th $\widehat{p}_{itcs}$'s that maximize $\mathcal{L}_n$; and $\widehat{\lambda} = n$. In short, the maximum likelihood fit is essentially the same for the fixed N and random N models, which is the main reason for the common choice `family = poisson` when fitting count data by maximum likelihood.

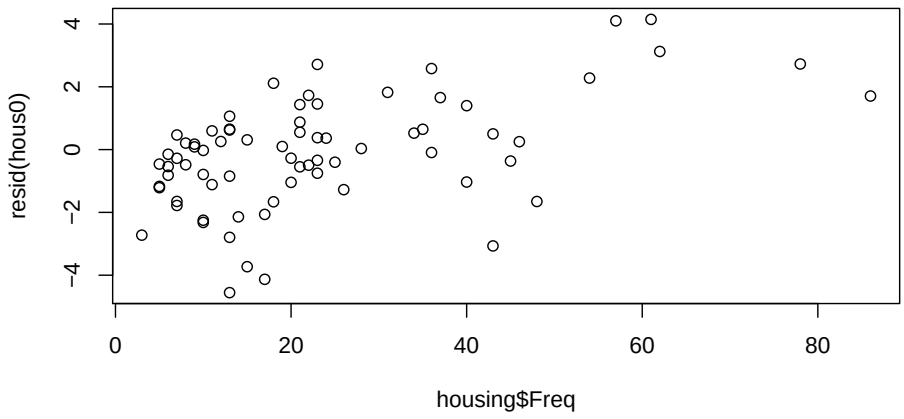
Let me try to reproduce the analysis given by (Venables and Ripley, 2002, Section 7.3). First they fitted the model

$$\log p_{itcs} = \theta_{itc} + \delta_s.$$

```
hous0 <- glm(Freq ~ Infl*Type*Cont + Sat,
    family = poisson, data = housing)
# for comparison with V\&R p200:
print(c(hous0$null.deviance,hous0$deviance, hous0$df.resid))

## [1] 833.657 217.456  46.000

plot(housing$Freq,resid(hous0))
```



Don't worry about the meaning of "deviance" for the moment. I included it just to check that I was fitting the same model as V&R. Not a great fit.

Next they fitted the model

$$\log p_{itcs} = \theta_{itc} + \theta_{si} + \theta_{st} + \theta_{sc}.$$

```
hous1 <- glm(Freq ~ Infl*Type*Cont  + Sat:(Infl+Type+Cont),
                 family = poisson, data = housing)
# for comparison with V\&R p200:
print(c(hous1$null.deviance,hous1$deviance,hous1$df.resid))

## [1] 833.6570  38.6622  34.0000

plot(housing$Freq,resid(hous1))
```

To conserve on space I'll abbreviate the factor levels before displaying the coefficients.

```
options(width=120)
is.Low <- grep("Low",housing$Sat)
Sat.matrix <- matrix(housing$Freq,byrow=T,ncol=3)
fit.matrix <- matrix(hous1$fit,byrow=T,ncol=3)
counts <- apply(Sat.matrix,1,sum)
satisfaction <- cbind(housing[is.Low,2:4],Sat.matrix,round(fit.matrix,1),
    round(sweep(Sat.matrix,1,counts,"/"),2),round(sweep(fit.matrix,1,counts,"/"),2))

names(satisfaction)[4:15] <- paste(rep(c("L", "M", "H"),times=4),
    rep(c("freq","est","prop","p.est"),each=3),sep=".")
levels(satisfaction$Type) <- c("Tw","Ap","At","Tr")
```

```
levels(satisfaction$Infl) <- c("L","M","H")
levels(satisfaction$Cont) <- c("L","M")
print(satisfaction)
```

```
##    Infl Type Cont L.freq M.freq H.freq L.est M.est H.est L.prop M.prop H.prop L.p.est M.p.est H.p.est
## 1     L   Tw    L     21     21     28  27.7  18.2  24.1   0.30   0.30   0.40    0.40    0.26    0.34
## 4     M   Tw    L     34     22     36  23.9  24.6  43.5   0.37   0.24   0.39    0.26    0.27    0.47
## 7     H   Tw    L     10     11     36   8.6  11.0  37.5   0.18   0.19   0.63    0.15    0.19    0.66
## 10    L   Ap    L     61     23     17  54.8  23.3  22.9   0.60   0.23   0.17    0.54    0.23    0.23
## 13    M   Ap    L     43     35     40  46.6  30.9  40.5   0.36   0.30   0.34    0.39    0.26    0.34
## 16    H   Ap    L     26     18     54  25.0  20.7  52.3   0.27   0.18   0.55    0.26    0.21    0.53
## 19    L   At    L     13      9     10  13.7  10.3   8.0   0.41   0.28   0.31    0.43    0.32    0.25
## 22    M   At    L      8      8     12   8.3   9.7  10.0   0.29   0.29   0.43    0.30    0.35    0.36
## 25    H   At    L      6      7      9   4.1   6.0  11.9   0.27   0.32   0.41    0.19    0.27    0.54
## 28    L   Tr    L     18      6      7  20.0   6.8   4.2   0.58   0.19   0.23    0.65    0.22    0.14
## 31    M   Tr    L     15     13     13  20.8  11.0   9.2   0.37   0.32   0.32    0.51    0.27    0.22
## 34    H   Tr    L      7      5     11   8.5   5.6   9.0   0.30   0.22   0.48    0.37    0.24    0.39
## 37    L   Tw    M     14     19     37  20.9  19.7  29.4   0.20   0.27   0.53    0.30    0.28    0.42
## 40    M   Tw    M     17     23     40  14.8  21.8  43.4   0.21   0.29   0.50    0.18    0.27    0.54
## 43    H   Tw    M      3      5     23   3.1   5.7  22.1   0.10   0.16   0.74    0.10    0.19    0.71
## 46    L   Ap    M     78     46     43  73.1  44.6  49.3   0.47   0.28   0.26    0.44    0.27    0.30
## 49    M   Ap    M     48     45     86  53.2  50.8  75.0   0.27   0.25   0.48    0.30    0.28    0.42
## 52    H   Ap    M     15     25     62  18.3  21.7  62.0   0.15   0.25   0.61    0.18    0.21    0.61
## 55    L   At    M     20     23     20  20.9  22.5  19.6   0.32   0.37   0.32    0.33    0.36    0.31
## 58    M   At    M     10     22     24  12.1  20.3  23.6   0.18   0.39   0.43    0.22    0.36    0.42
## 61    H   At    M      7     10     21   4.9  10.2  22.9   0.18   0.26   0.55    0.13    0.27    0.60
## 64    L   Tr    M     57     23     13  50.9  24.6  17.5   0.61   0.25   0.14    0.55    0.27    0.19
## 67    M   Tr    M     31     21     13  26.3  19.9  18.8   0.48   0.32   0.20    0.40    0.31    0.29
## 70    H   Tr    M      5      6     13   6.6   6.2  11.3   0.21   0.25   0.54    0.27    0.26    0.47
```

The last 12 columns of satisfaction denote

$$\begin{aligned} \text{[HML].freq} &= \text{observed counts} \\ \text{[HML].est} &= \text{estimated counts} \\ \text{[HML].prop} &= \text{observed proportions} \\ \text{[HML].p.est} &= \text{estimated probabilities} \end{aligned}$$

Venables and Ripley (2002, page 202) commented:

The message of the fitted model is now clear. The factor having most effect on the probabilities is influence, with an increase in influence reducing the probability of low satisfaction and increasing that of high. The next most important factor is the type of housing itself. Finally, as contact with other residents rises, the probability of low satisfaction tends to fall and that of high to rise, but the effect is relatively small. \\ *The reader should compare the model-based probability estimates with the relative frequencies from the original data. In a few cases the smoothing effect of the model is perhaps a little larger than might have been anticipated, but there are no very surprising differences.*

Then V&R started to fiddle around with other slightly different models. Did we learn anything new by fitting the `hous1` model?

2 Fitting

For general Poisson log-linear models we have counts y_α, for $\alpha \in \mathbb{A}$ ranging over all the cells in the cross-tabulation for p factors. These y_α's are modelled as independent Poisson(μ_α) random variables with $\log \mu_\alpha = \theta_\alpha = w_\alpha^T b$, for a specified set of factor levels w_α for cell α. If there are N cells in the table then the linear predictor for the vector θ equals Xb, where X is the $N \times p$ matrix with w_α^T as its αth row.

Remark. I have indexed the set of cells by $\mathbb{A}$ instead of the usual $\{1, \ldots, N\}$ to avoid confusing myself when thinking of $\mathbb{A}$ as the set of all combinations of levels for the factors.

We can think of the likelihood as a function of μ or of θ or, if $\theta = Xb$, as a function of b. Even though the last parametrization might seem the most appropriate, it is worthwhile to consider values of θ that do not lie in span(X). You will see soon the reason for this relaxation of the rules. The log-likelihood equals

$$\begin{aligned}\mathcal{L}(\theta) &= \sum\nolimits_\alpha g_\alpha(\theta_\alpha) \\ &\qquad \text{where } g_\alpha(\theta) = \log\left(e^{-\mu_a}\mu_a^{y_\alpha}/y_\alpha!\right) = y_\alpha\theta_\alpha - e^{\theta_\alpha} - \log(y_\alpha!).\end{aligned}$$

Note that $g_a(t)$ has first and second derivatives (with respect to θ_α)

$$\begin{aligned}\dot{g}_\alpha(\theta_\alpha) &= y_\alpha - e^{\theta_\alpha} = y_\alpha - \mu_a \\ \ddot{g}_\alpha(\theta_\alpha) &= -e^{\theta_\alpha} = -\mu_a.\end{aligned}$$

Consider the effect of making a (small?) change in θ. By Taylor expansion,

$$\begin{aligned}\mathcal{L}(\theta + h) &= \sum_\alpha g_\alpha(\theta_\alpha + h_\alpha) \\ &\approx \sum\nolimits_\alpha \left(g_\alpha(\theta_\alpha) + h_\alpha\dot{g}_\alpha(\theta_\alpha) + \tfrac{1}{2}h_\alpha^2\ddot{g}_\alpha(\theta_\alpha)\right) \\ &= \mathcal{L}(\theta) + \tfrac{1}{2}\sum\nolimits_\alpha \left(h_\alpha(y_\alpha - \mu_\alpha) - \tfrac{1}{2}h_\alpha^2\mu_\alpha\right) \\ &= \mathcal{L}(\theta) + \tfrac{1}{2}\sum\nolimits_\alpha (y_\alpha - \mu_\alpha)^2/\mu_\alpha - \tfrac{1}{2}\sum\nolimits_\alpha \mu_\alpha\left[(y_\alpha - \mu_\alpha)/\mu_\alpha - h_\alpha\right]^2. \qquad \text{<1>}\end{aligned}$$

The last equality comes from completing the square. Define

$$\begin{aligned}\eta &= \theta + h \\ z_\alpha = z_\alpha(\theta) &= \theta_\alpha + (y_\alpha - \mu_\alpha)/\mu_\alpha \\ Q(\mu) &= \sum\nolimits_\alpha (y_\alpha - \mu_\alpha)^2/\mu_\alpha.\end{aligned}$$

Then the approximation can be rewritten as

<2> $$\mathcal{L}(\eta) \approx \mathcal{L}(\theta) + \tfrac{1}{2}Q(\mu) - \tfrac{1}{2}\sum\nolimits_\alpha \mu_\alpha \left(z_\alpha - \eta_\alpha\right)^2 .$$

This approximation suggests an algorithm for determining the $\widehat{b}$ that maximizes $\mathcal{L}(Xb)$ over b in $\mathbb{R}^p$.

(i) Initialize by choosing μ equal to y, that is, $\theta_\alpha = \log y_\alpha$.

(ii) Iterate until the θ "seems to have converged":

(a) Define $\mu_\alpha = \exp(\theta_\alpha)$ and $z_\alpha = \theta_\alpha + (y_\alpha - \mu_\alpha)/\mu_\alpha$.

(b) Find b to minimize $\sum_\alpha \mu_\alpha \left(z_\alpha - \eta_\alpha\right)^2$ where $\eta = Xb$.

(c) Redefine: $\theta_\alpha = \eta_\alpha$.

It doesn't matter that the θ for the initialization step (i) might not be of the form Xb. We just have to take some precaution against cases where some y_α is zero. The derivation of <2> still makes some sense if we don't worry about h being small.

The minimization in step (ii) is a weighted least squares problem. If we write M for $\operatorname{diag}(\sqrt{\mu}_\alpha : \alpha \in \mathbb{A})$ then the problem can be written as

$$\text{find } b \text{ to minimize } (z - Xb)^T M^2 (z - Xb) = \|Mz - MXb\|_2^2 .$$

Remark. **R** handles weighted least squares by using the function `lm.wfit()`. In the code for that function there are the lines

```
wts <- sqrt(w)
z <- .Call(C_Cdqrls, x * wts, y * wts, tol, FALSE)
```

That is, it uses the ordinary `qr` method after multiplying the response and the predictors by the square root of the weights.

Somewhere in that function, or in `glm()`, **R** must be adding in the intercept term.

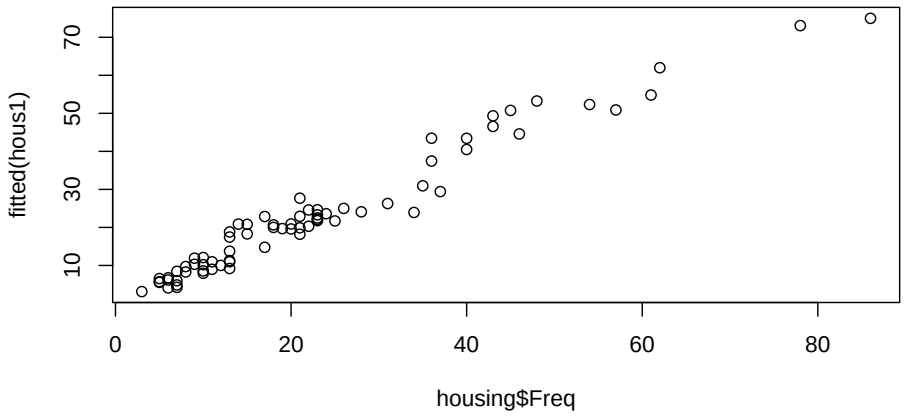
3 Generalized linear models

I was going to point out that the log-linear Poisson model fits into the generalized framework described by (McCullagh and Nelder, 1989, Section 2.5). Instead let me point you to an old handout, GLM2010.pdf.

4 Deviance and measures of fit

The housing data set contains counts of numbers of individuals (`Freq`) for each of the combinations of the four factors. The models led to estimated counts. We need some measure of how close these two sets of counts are to each other.

```
plot(housing$Freq,fitted(hous1))
```



Remember that

$$\log \mathcal{L}(\theta) = \sum\nolimits_{\alpha\in\mathbb{A}} \Big(y_\alpha \theta_\alpha - e^{\theta_\alpha} - \log(y_\alpha!) \Big)$$

There are several common ways to measure how close the fitted values are to the data. Suppose $\widehat{\theta}$ maximizes $\mathcal{L}(\theta)$ subject to the constraint that $\widehat{\theta} \in \text{span}(X)$. That is, $\widehat{\theta} = X\widehat{b}$ for the $\widehat{b}$ that maximizes $\mathcal{L}(Xb)$. If there were no constraint on the θ then elementary calculus shows that the maximum is achieved at $\theta^*_\alpha = \log y_\alpha$. The quantity

$$D_X = 2\mathcal{L}(\theta^*) - 2\mathcal{L}(\widehat{\theta})$$

is called the deviance for the model.

More classical is the chi-squared statistic

$$\sum\nolimits_\alpha (y_\alpha - \widehat{\mu}_\alpha)^2/\widehat{\mu}_\alpha \qquad \text{where } \widehat{\mu}_\alpha = e^{\widehat{\theta}_\alpha}$$

or its modified form

$$\sum\nolimits_\alpha (y_\alpha - \widehat{\mu}_\alpha)^2/y_\alpha.$$

The last two statistics are closely related to the squared distance between square roots,

$$\sum\nolimits_\alpha \left(\sqrt{y}_\alpha - \sqrt{\widehat{\mu}}_a\right)^2 = \sum\nolimits_\alpha \frac{(y_\alpha - \widehat{\mu}_a)^2}{\left(\sqrt{y}_\alpha + \sqrt{\widehat{\mu}}_a\right)^2},$$

which is suggested by the fact that $\sqrt{y_\alpha}$ is approximately $N(\sqrt{\lambda}_\alpha, 1/4)$ distributed if $y_\alpha \sim \text{Poisson}(\lambda_\alpha)$.

In an asymptotic sense all of these measures of fit are capturing the same idea.

4.1 Asymptotics

Suppose the y_α's are actually independent, with $y_\alpha \sim \text{Poisson}(\lambda_\alpha)$ and $\log \lambda_\alpha = w_\alpha^T \beta$ for some unknown β in $\mathbb{R}^p$. Define $m_\alpha = \sqrt{\lambda}_\alpha$ and $\xi_\alpha = (y_\alpha - \lambda_\alpha)/m_\alpha$ and $M = \text{diag}(m_\alpha : \alpha \in \mathbb{A})$. Then approximation <1> can be rewritten as

$$\begin{aligned}\mathcal{L}(X\beta + h) &\approx \mathcal{L}(X\beta) + \tfrac{1}{2}\sum\nolimits_\alpha \left(h_\alpha(y_\alpha - \lambda_\alpha) - \tfrac{1}{2}h_\alpha^2\lambda_\alpha\right) \\ &= \mathcal{L}(X\beta) + \tfrac{1}{2}\sum\nolimits_\alpha \xi_\alpha^2 - \tfrac{1}{2}\sum\nolimits_\alpha (\xi_\alpha - m_\alpha h_\alpha)^2.\end{aligned}$$

For a general b in $\mathbb{R}^p$ define $t = b - \beta$ and $h = Xb - X\beta$. Then the last approximation becomes

$$\mathcal{L}(Xb) = \mathcal{L}(X\beta + h) \approx \mathcal{L}(X\beta) + \tfrac{1}{2}\left\|\xi\right\|^2 - \tfrac{1}{2}\left\|\xi - MXt\right\|^2.$$

The $\widehat{b}$ that maximizes the left-hand side corresponds (approximately) to the $\widehat{t}$ that minimizes $\|\xi - MXt\|^2$.

Once again we have a least squares problem. If $\mathcal{H}$ denotes the matrix that projects orthogonally onto $\text{span}(MX)$ then

$$\mathcal{L}(X\widehat{b}) \approx \mathcal{L}(X\beta) + \tfrac{1}{2}\left\|(I - \mathcal{H})\xi\right\|^2.$$

By the central limit theorem the vector ξ is approximately $N(0, I_N)$-distributed. The term $\|(I - \mathcal{H})\xi\|^2$ is approximately χ^2-distributed.

If I were to pursue this idea further you would see how other goodness-of-fit quantities involve terms that are approximately χ^2-distributed, which leads to (approximate) ways to compare the fits for various models.

5 Structural zeros

Read McCullagh and Nelder (1989, Section 3.7).

References

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (4th ed.). Springer-Verlag.