Count data

To be continued.

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

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

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 $\log \mathcal{L}_n$:

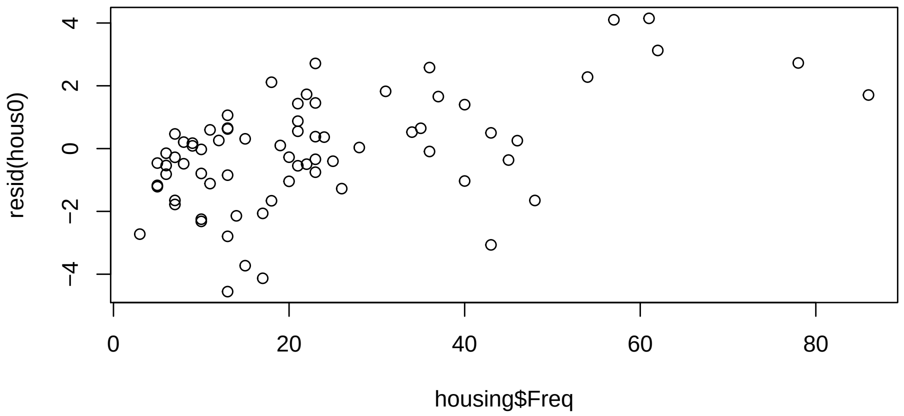
$$\text{log-likelihood} = -\lambda + n \log \lambda - \log(n!) + \mathcal{L}_n.$$

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))
```

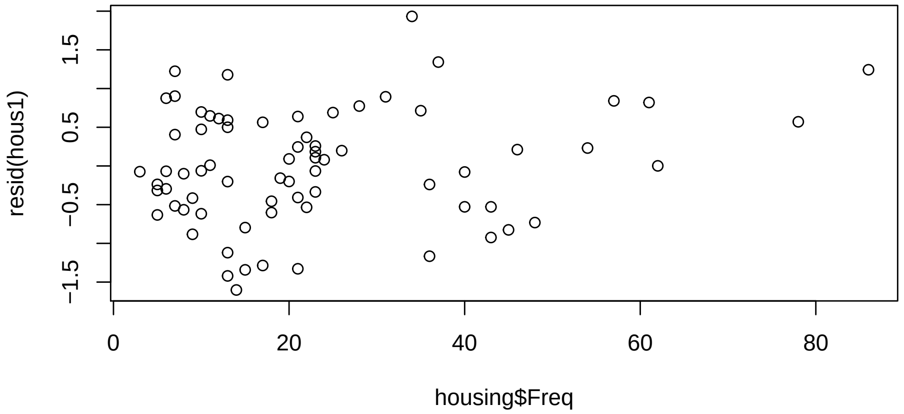


Don’t worry about the meaning of “deviance” for the moment. I included it just to check that I was fitting the same model as V&R. Not a great fit.

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.

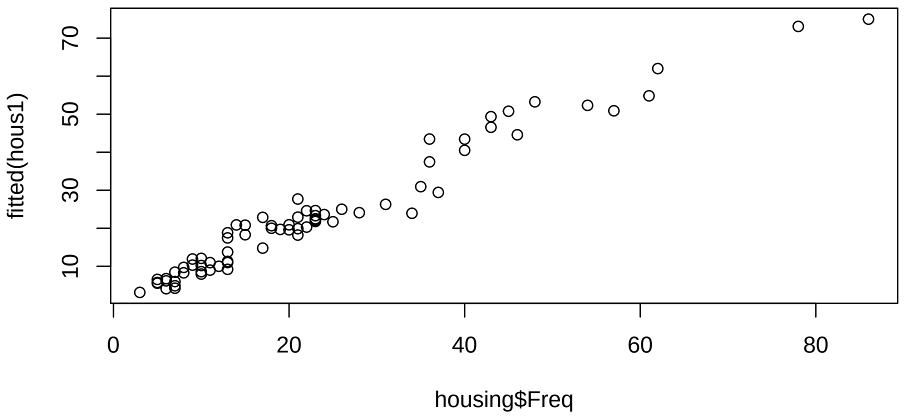
```
 round(hous1.abb$coeff,1)

## (Intercept)          iM          iH         tAp         tAt         tTr
##         3.1         0.2        -0.4         0.3        -0.8        -1.0
##          cM      iM:tAp      iH:tAp      iM:tAt      iH:tAt      iM:tTr
##         0.0         0.0         0.4        -0.4         0.0         0.2
##      iH:tTr       iM:cM       iH:cM      tAp:cM      tAt:cM      tTr:cM
##          0.3        -0.2        -0.7         0.6         0.7         1.2
##       iL:s.L      iM:s.L      iH:s.L      iL:s.Q      iM:s.Q      iH:s.Q
##         -0.1         0.4         1.0         0.3         0.2         0.4
##      tAp:s.L     tAt:s.L     tTr:s.L     tAp:s.Q     tAt:s.Q     tTr:s.Q
##         -0.5        -0.3        -1.0         0.1        -0.3         0.0
##       cM:s.L      cM:s.Q  iM:tAp:cM  iH:tAp:cM  iM:tAt:cM  iH:tAt:cM
##          0.3        -0.1         0.0         0.1         0.2         0.5
##    iM:tTr:cM   iH:tTr:cM
##         -0.5        -0.5
```

Then V&R started to fiddle around with other slightly different models.

2 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 lead to estimated counts. We need some measure of how close these two sets of counts are to each other.

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



There are several common ways to measure how close the fitted values are to the data. [To be continued.]

3 Structural zeros

(McCullagh and Nelder, 1989, page 14; Chapter 6)

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.