**Log-Linear Modelling**

One application of Poisson models - when $Y$ is clearly a "count" variable.

For another application, consider the "housing" data (Sec. 7.3, text).

Survey of $n = 1681$ renters in Copenhagen, asked:
**satisfaction** with housing condition (L, M, H),
**type** of housing (tower block, apartment, atrium, terrace),
degree of **contact** with neighbours (L, H),
**influence** on management (L, M, H).

Interest in how $Sat$ is explained by $(Type, Infl, Cont)$.
GLM?



$$\{Pr(Sat = L), Pr(Sat = M), Pr(Sat = H)\} \quad = \quad (p_1, p_2, 1 - p_1 - p_2).$$

i.e., **multinomial** response.

Would need 2-D link function:

$$g(p_{1i}, p_{2i}) \quad = \quad \beta_1 X_{1i} + \ldots + \beta_p X_{pi},$$

for $i = 1, \ldots, 1681$.



Or, take a different view of the data structure: responses are **frequencies** associated with all possible combinations of levels for $(SAT, TYPE, INFL, CONT)$.

```
> 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
...
68 Medium Medium   Terrace High   21
69   High Medium   Terrace High   13
70    Low   High   Terrace High    5
71 Medium   High   Terrace High    6
72   High   High   Terrace High   13
```



Poisson GLM (with log-link) to explain $FREQ$ in terms of $(SAT, INFL, TYPE, CONT)$?

Bearing in mind the real interest is in explaining $Sat$ in terms of $(Infl, Type, Cont)$.

**CLAIM:** smallest interesting/relevant/appropriate model is

$$FREQ \quad \sim \quad INFL * TYPE * CONT + SAT,$$

as this corresponds to $(Sat|Infl, Type, Cont)$ *not* depending on $(Infl, Type, Cont)$ (like an intercept-only model).

Then start model-building by adding interactions, e.g.,

$$FREQ \sim INFL * TYPE * CONT + SAT + SAT : CONT$$

This corresponds to $(Sat|Infl, Type, Cont)$ depending on $Type$, but not on $(Infl, Cont)$.

Another ex.,

$$\begin{aligned} FREQ \sim\ & INFL * TYPE * CONT + SAT + SAT : CONT + \\ & SAT : TYPE + SAT : INFL + SAT : TYPE : CONT \end{aligned}$$

would correspond with

$$Sat \sim Cont + Type + Infl + Type : Cont$$



Why should/must the INFL*TYPE*CONT terms be included, i.e., why is part of the model necessarily saturated?

Has the desirable property that the fitted values = observed frequencies for the (INFL, TYPE, CONT) “marginal,” i.e., think of summing fitted and actual frequencies over the $SAT$ variable.

**Interpretation:** we aren’t doing any modelling for the distribution of the predictors, only for the distribution of the response given the predictors.



In fact, multinomial modelling for a categorical response variable (and categorial predictors) can be shown to be mathematically equivalent to Poisson modelling for the corresponding frequencies (see text for an empirical example of this).

The idea of ‘saturating’ part of the Poisson model to reflect relationships one isn’t trying to model is quite common. A related idea is that sometimes some margins are ‘fixed by design,’ and the corresponding fitted values better match.