

# Counts and Proportions: Logistic and Log-linear models

# Observing Count Data

Statistics is often about

‘what might have happened but did not’

and for count data that sometimes matters. Suppose we have a  $2 \times 2$  table of counts, and we think of the rows as an explanatory factor and the columns as a response factor.

	resp1	resp2	Totals
cause1	$n_{11}$	$n_{12}$	$n_{1.}$
cause2	$n_{21}$	$n_{22}$	$n_{2.}$
Totals	$n_{.1}$	$n_{.2}$	$n$

This could have arisen in many different ways.

# Sampling Schemes for $2 \times 2$ Tables

- *Poisson*. We collect data for a certain period of time, and  $n$  and all  $n_{ij}$  are Poisson-distributed, the latter independently for all pairs  $i, j$ .
- *multinomial*, in which  $n$  is fixed and the numbers in the cells are from a multinomial distribution.
- *product multinomial*. Here the  $n_{i\cdot}$  are fixed in advance: we pick fixed numbers for each level of the explanatory factor and have an independent multinomial (binomial for two response levels) distribution in each row. So the row totals are fixed.
- *hypergeometric*. Here all the row and column totals are fixed. Suppose a woman is given 14 cups of tea and told 7 had milk poured in first and 7 has tea poured in first and is asked to identify them. Unusual!
- *retrospective product binomial*. Here the column totals are fixed, as the numbers of the two response categories (often called *controls* and *cases*) are fixed and the sampling generates different numbers of each cause.

# Conditioning

We can often turn one sampling scheme into another by conditioning on ancillary statistics. For example, Poisson sampling conditioned on the total number is multinomial sampling, and each of Poisson and multinomial sampling conditioned on the row totals is product multinomial sampling, or conditioned on the column totals is retrospective multinomial sampling.

The key thing is not to attempt inference on something that was fixed by the sampling scheme – special care is needed in retrospective sampling. Also watch for any additional restrictions: *matched* case–control studies are not covered by the methods here, and need additional conditioning.

# Differences in Proportions

Under product multinomial sampling we have probabilities

	resp1	resp2
cause1	$1 - \pi_1$	$\pi_1$
cause2	$1 - \pi_2$	$\pi_2$

and the simplest question is to ask about  $\pi_2 - \pi_1$ , in particular if it is zero. It is obvious that we should use the MLE  $\hat{\pi}_i = n_{i2}/n_{i.}$ , and also that  $n_{i2}$  has a binomial distribution, which gives the mean and variance of  $\hat{\pi}_2 - \hat{\pi}_1$ . Using the CLT we can get approximate tests of equality and confidence intervals for  $\pi_2 - \pi_1$ .

However, it is normally better to use *odds* than probabilities:

define  $\omega_i = \pi_i/(1 - \pi_i)$ .

In gambling circles one has odds like ‘10 to 1’ and ‘2 to 1 on’ (really 1 to 2) which suggest odds are naturally on a log scale.

# Odd Ratios

The most important parameter in the analysis of a  $2 \times 2$  table is the odds ratio  $\phi = \omega_2/\omega_1$ : an odds ratio of one implies the probabilities are equal (and hence no association in the table). The obvious estimate of the odds ratio is

$$\hat{\phi} = \frac{\hat{\pi}_2/(1 - \hat{\pi}_2)}{\hat{\pi}_1/(1 - \hat{\pi}_1)} = \frac{n_{22}/n_{21}}{n_{12}/n_{11}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

sometimes known as the *cross-product ratio* or  $ad/bc$  if the cells counts are labelled  $a$  to  $d$ .

We can re-express the test of  $\pi_2 = \pi_1$  as a test for  $\phi = 1$ , and develop a confidence interval for  $\log \phi$  based on  $\log \hat{\phi}$ .

The odds ratio is the same whichever factor is thought of as the response, so inference for the odds ratio is valid for retrospective (case-control) studies. It can also be done for hypergeometric sampling.

## An example

The US General Social Surveys in 1982 and 1994 asked respondents to agree or disagree with the statement

‘Women should take care of running their homes and leave running the country up to men’

The numbers of responses were

Year	disagree	agree	total
1982	223	122	345
1994	1632	268	1900

This is presumably product multinomial sampling (given the round number). Clearly  $\hat{\pi}_1 = 122/345 \approx 0.35$  and  $\hat{\pi}_2 = 268/1900 \approx 0.14$ . Is this a significant difference? A 95% confidence interval for  $\pi_2 - \pi_1$  is  $(-0.26, -0.16)$ . On the other hand, the odds ratio is  $268 \times 223 / 122 \times 1632 = 0.30 = 1/3.33$  with confidence interval  $(0.23, 0.39)$ .

## Mantel–Haenszel Test

Suppose we have several  $2 \times 2$  tables, perhaps different studies of the same effect. The Mantel–Haenszel test is a test of odds-ratio one in each of the tables, *assuming that they all have the same odds ratio*. This works for product multinomial or retrospective sampling.

A classic example tested the effectiveness of immediately injected or 1.5 hours delayed penicillin in protecting rabbits against a lethal injection with beta-hemolytic streptococci. There was one table for each of five dose levels of penicillin. The  $P$  value was about 4%, indicating some extra effectiveness of immediate injection although none of the tables individually were significant. The 95% confidence interval for the common odds ratio was (1.03, 47).

Such problems can be investigated much more carefully by *logistic regression*.

# Logistic Regression

There were many lawsuits in the US by 1999 over the safety of tyres of the Ford Explorer, a ‘compact’ SUV. For 2,321 fatal accidents involving such vehicles the cause (tyre failure or not) and the make of vehicle (condensed to Ford or not) were recorded.

	Non-tyre	Tyre
Ford	500	22
Other	1974	5

which has a chi-squared value of 51.2 and a negligible  $P$  value.

However, we know much more about each accident, for example the age of the vehicle and the number of passengers. We *could* stratify on these variables and do a Mantel–Haenszel test, if we believed in a common odds ratio.

## Logistic Regression 2

For each accident we have a binary response. Let  $p$  denote the probability of the cause being a tyre failure (a ‘success’) and consider the log-odds of success,

$$\text{logit}(p) = \log \frac{p}{1-p} = \eta$$

say. Then

$$p = \frac{e^\eta}{1 + e^\eta}$$

the *logistic* function.

In logistic regression we postulate that  $\eta$  is a linear function of explanatory variables

$$\text{logit}(p) = \eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

## Logistic Regression 3

Consider first a single binary regressor  $x_1 \in \{0, 1\}$  and corresponding probabilities  $p_0$  and  $p_1$ . Then

$$\beta_1 = \text{logit}(p_1) - \text{logit}(p_0) = \log \frac{p_1(1 - p_0)}{p_0(1 - p_1)} = \log \phi$$

the log of the odds-ratio  $\phi$  between groups 1 and 0. So we can interpret coefficients via odds ratios.

### Probit models

An alternative is to replace logit by *probit*, the quantile function for the normal rather than the logistic distribution.

$$\Phi^{-1}(p) = \eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Normally very similar except a scale factor of about 1.7 on the  $\beta_i$ .

# Maximum Likelihood Fitting

We have 2,321 accidents, and for each the outcome is  $y_i = 0$  (failure) or 1 success. The log-likelihood is

$$L(\beta) = \sum_{i:y_i=1} \log p_i + \sum_{i:y_i=0} \log(1 - p_i)$$

where  $p_i$  depends on  $\eta_i$  which depends on  $\beta$ . This is negative, and often we reverse the sign.

Note that  $-L$  is a figure of how well we predicted sometimes known as *log-probability scoring*.  $-2L$  is known as the *deviance* (in this case: see later).

In general  $-L$  has to be minimized numerically, and it is possible that it does not have a minimum (it decreases as  $\beta$  goes to infinity in some direction, and computer programs may give non-convergence messages).

# Grouped vs Ungrouped Data

There are 240 different possible combinations of make, cause, age and number of passengers (and only 86 exist in the dataset), and 2,321 accidents (27 from tyres). So we could group the data by saying that for each combination  $x_j$  of explanatory variables there were  $Y_j$  out of  $n_j$  successes (accidents caused by tyre failure). Now the likelihood will be different, with each group  $j$  having  $Y_j \sim \text{Binomial}(n_j, p_j)$ .

Fact: the maximum likelihood estimates will be the same as for the ungrouped data, and the increase in log-likelihood when we add extra regressors will also be the same. For a given model and  $\beta$  the two log-likelihoods differ only by an additive constant.

## Social Attitudes again

We can fit the social attitudes table via logistic regression.

```
> att <- data.frame(disagree=c(223,1623), agree=c(122,268),  
                    time=0:1, row.names=c(1984, 1992))  
> summary(glm(cbind(agree, disagree) ~ time, data=att,  
              family=binomial), cor = F)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.603	0.113	-5.36	8.5e-08
time	-1.198	0.130	-9.18	< 2e-16

so the odds ratio is estimated as  $\exp -1.198 \approx 0.302$ , with a 95% confidence interval of  $\exp(-1.198 \pm 1.96 \times 0.130) = (0.234, 0.389)$ .

# Deviance for Grouped Data

A more complete definition of the deviance is

$$2 \left( L(\text{saturated model}) - L(\hat{\beta}) \right)$$

where ‘saturated model’ means a model with the best possible fit, that is a parameter for each group. (For ungrouped data there is a parameter for each observation and so the fitted probability can be 0 or 1 for observed success or failure and so the first term is zero. For grouped data we have  $\hat{p}_j = Y_j/n_j$ .)

For grouped data the residual deviance can be a useful measure of lack of fit, and if we fitted  $p$  parameters to  $M$  groups it has approximately a  $\chi^2_{M-p}$  distribution. For a single categorical regressor this gives the  $G^2 = \sum E_j \log O_j/E_j$  test (summed over numbers of successes and of failures), closely related to the chi-squared test. The approximation is poor if the expected numbers of either successes or failures are small.

# Analysis of Deviance

We can treat the deviance in almost the same way as the residual sum of squares, with the analysis of deviance replacing the analysis of variance: the only practical difference is that we assume we know  $\sigma^2 = 1$  and so chi-squared distributions are used in place of  $F$  distributions.

This means that ‘drop-in-deviance’ tests can be used to test if terms should be added or dropped from a logistic regression: they are just twice logs of likelihood ratio test statistics. Remember that the distribution theory is relying on large-sample theory.

# Inference for Coefficients

From standard asymptotic distribution theory, standard errors for the components of  $\hat{\beta}$  can be computed, and from those  $t$  ratios. Resist the temptation to assume that the  $t$  ratios are  $t$ -distributed. They are approximately normally distributed, and provide Wald tests of the non-zerosness of individual coefficients.

**However**, there is a fatal flaw. As pointed out by Hauck & Donner (1977), when  $\hat{\beta}_j$  is large, so is its estimated standard deviation and so the  $t$  ratio is small. Thus a small  $t$  ratio indicates that **either** a coefficient is not significantly different from zero **or** that it is very significantly different, but not which. A drop-in-deviance test does not have this disadvantage, and sometimes the actual size of  $\hat{\beta}_j$  can suggest which.

Since this means that confidence intervals based on the  $t$  ratios will be unreliable: it is better to base them on profile likelihood plots.

# SUV Accidents

First we group the data: working with 27/2321 ‘successes’ is perilous.

```
attach(ex2018)
ind <- vehicle.age + 5*passengers + 100*(make=="Ford")
tab <- as.data.frame(t(table(cause,ind)))
id <- as.numeric(row.names(tab))
tab$make <- id%%100
id <- id%%100
tab$passengers <- id %% 5
tab$vehicle.age <- id %% 5
names(tab)[1:2] <- c("Other", "Tyre")
detach()
```

```
> options(contrasts=c("contr.treatment", "contr.poly"))
> fit <- glm(cbind(Tyre, Other) ~ vehicle.age + passengers + make,
             data=tab, family=binomial)
> summary(fit, cor=F)
```

Coefficients:

	Value	Std. Error	t value
(Intercept)	8.14276	0.74735	10.8955
vehicle.age	0.48090	0.17330	2.7749
passengers	0.63151	0.10045	6.2866
make	2.77755	0.51827	5.3593

Null Deviance: 162.07 on 70 degrees of freedom

Residual Deviance: 78.652 on 67 degrees of freedom

This is linear in vehicle age and number of passengers. This is coded so Ford has a higher risk of tyre accidents with log odds ratio 2.8 .

We can do better.

```
> fit <- glm(cbind(Tyre, Other) ~
             factor(vehicle.age) + factor(passengers) + make,
             data=tab, family=binomial)
> summary(fit, cor=F)
```

Coefficients:

	Value	Std. Error	t value
(Intercept)	7.420282	0.99893	7.4282556
factor(vehicle.age)1	-2.874313	1.15074	-2.4978014
factor(vehicle.age)2	-2.094639	0.92504	-2.2643693
factor(vehicle.age)3	0.525048	0.66336	0.7915001
factor(vehicle.age)4	0.052693	0.70919	0.0743006
factor(passengers)1	0.648534	0.92604	0.7003280
factor(passengers)2	2.767846	0.84522	3.2747140
factor(passengers)3	2.987613	0.85017	3.5141422
factor(passengers)4	4.083281	0.88486	4.6145878
factor(passengers)5	3.857509	1.07455	3.5898876
factor(passengers)6	4.563173	1.45395	3.1384597
factor(passengers)7	-1.800245	30.07543	-0.0598577

```
factor(passengers)9 -2.757913    60.43962 -0.0456309
factor(passengers)11  0.312185    60.44171  0.0051651
                make  3.070097     0.53603  5.7275011
```

Null Deviance: 162.07 on 70 degrees of freedom

Residual Deviance: 42.717 on 56 degrees of freedom

The counts here are very small, so we need to be careful not to over-interpret the residual deviances as a very good fit.

```
> dropterm(fit, test="Chisq")
```

Single term deletions

Model:

```
cbind(Tyre, Other) ~ factor(vehicle.age) + factor(passengers) + make
```

	Df	Deviance	AIC	LRT	Pr(Chi)
<none>		42.716	72.72		
factor(vehicle.age)	4	72.791	94.79	30.074	4.7266e-06
factor(passengers)	9	91.538	103.54	48.821	1.7930e-07
make	1	88.694	116.69	45.977	0.0000e+00

So all the terms are significant, and the fit is good unlike

```
> glm(cbind(Tyre, Other) ~ make, data=tab, family=binomial)
```

Coefficients:

```
(Intercept)  make  
5.8826  2.759
```

Degrees of Freedom: 71 Total; 69 Residual

Residual Deviance: 119.11

# Residuals and Diagnostics

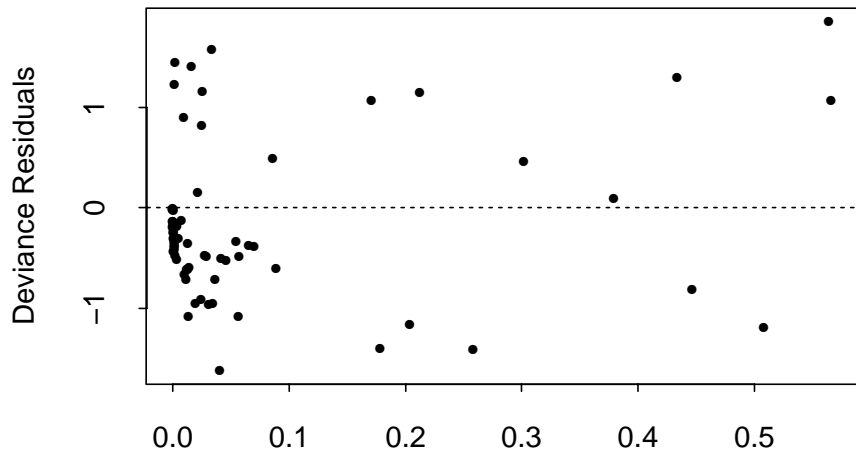
Using residuals to examine fits is much harder than in linear regression, mainly due to the discreteness of the response. There are various types of residuals for logistic regressions, including

**response** residuals  $Y_j - \hat{p}_j$  — not much use for binary data.

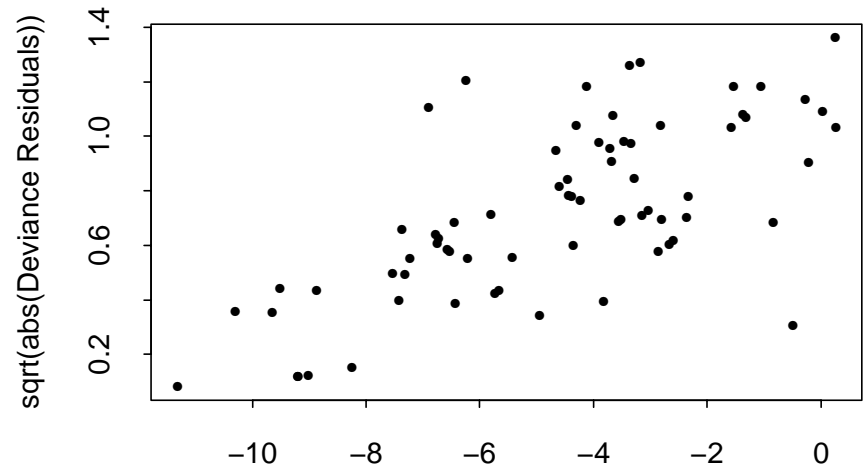
**deviance** residuals — the signed square root of the contribution to the deviance for that group.

**Pearson** residuals — the signed square root of the contribution to the chi-squared statistic for that group.

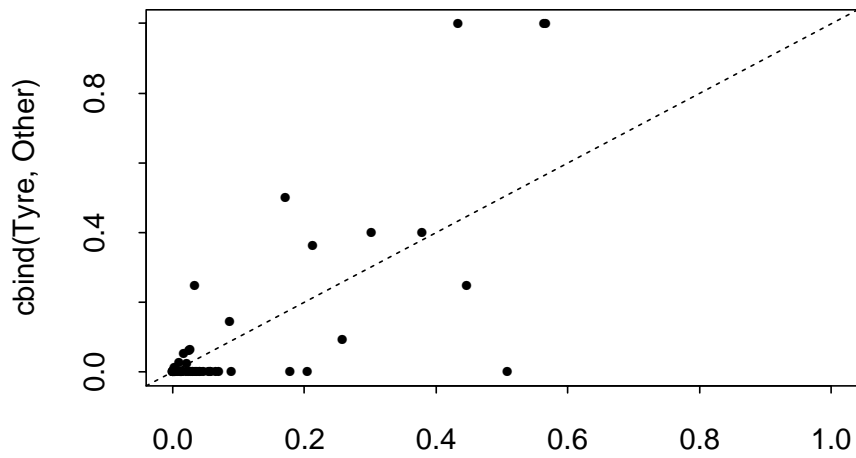
Diagnostic plots can be made but are generally hard to interpret. High leverage is much less of a problem as the observations are bounded.



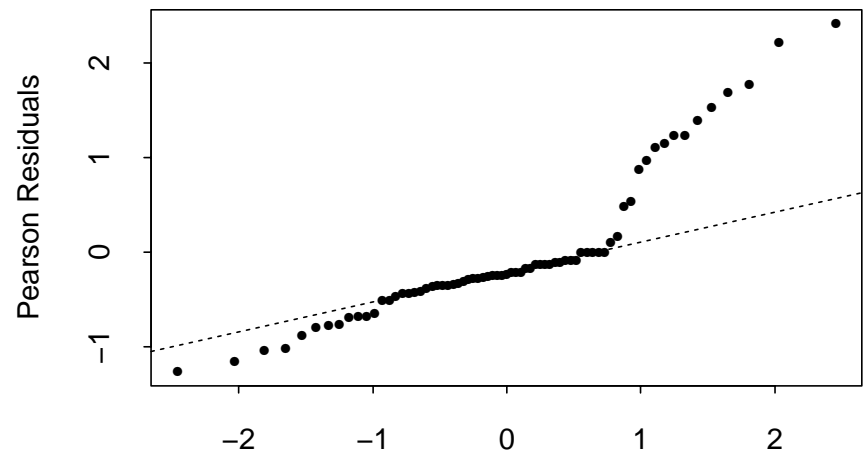
Fitted : factor(vehicle.age) + factor(passengers) + make



Predicted : factor(vehicle.age) + factor(passengers) + make

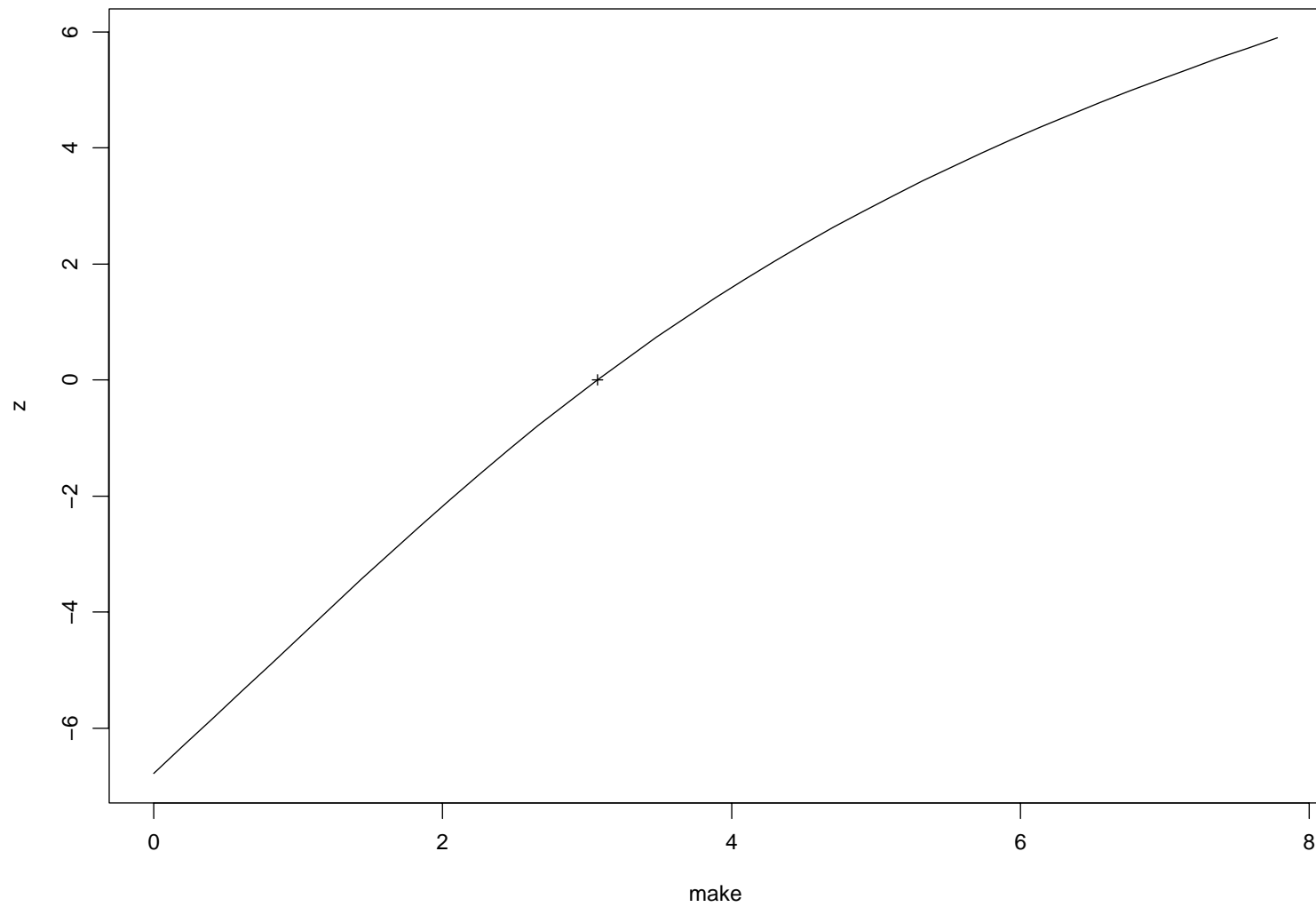


Fitted : factor(vehicle.age) + factor(passengers) + make



Quantiles of Standard Normal

# Signed Square-root Profile Likelihood Plot

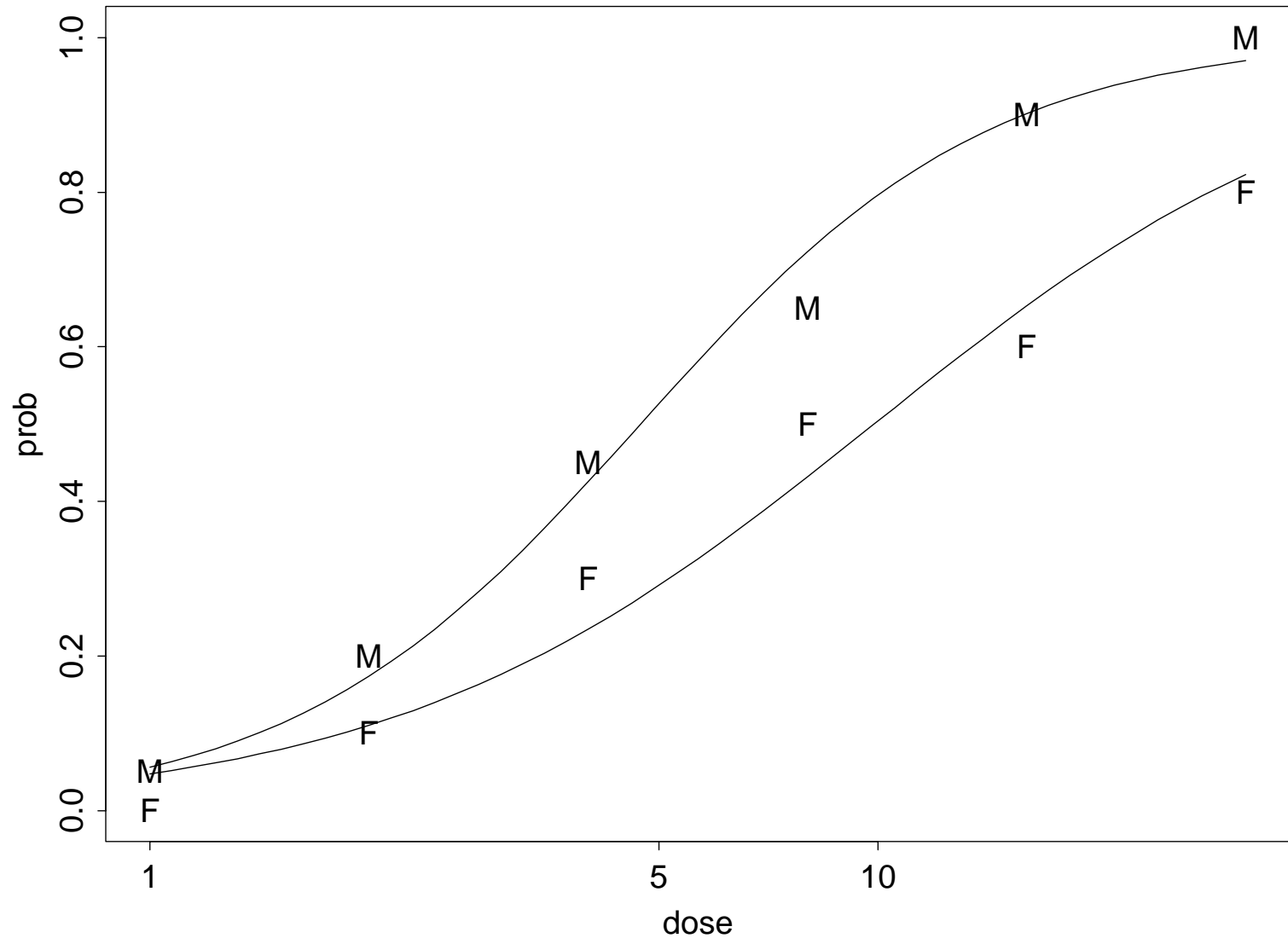


## ED50 (or LD50)

Collett (1991) reports an experiment on the toxicity to the tobacco budworm *Heliothis virescens* of doses of the *trans*-cypermethrin to which the moths were beginning to show resistance. Batches of 20 moths of each sex were exposed for three days to the pyrethroid and the number in each batch that were dead or knocked down was recorded. The results were

	Dose in $\mu\text{g}$					
Sex	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

We fit a logistic regression model using  $\log_2(\text{dose})$  since the doses are powers of two. The chosen model has parallel lines for the sexes on logit scale.



The main interest is in the dose at which 50% of the budworms die. We need to find the value of  $x$  for which  $\text{logit}(p) = 0$ . We can find this and, using the delta method, its standard error. For females we find

```
> dose.p(budworm.lg0, cf = c(1,3), p = 1:3/4)
      Log Dose      SE
p = 0.25: 2.2313 0.24983
p = 0.50: 3.2636 0.22971
p = 0.75: 4.2959 0.27462
```

and lower values for the males. Note that we do this on log scale as we would want e.g. to form confidence intervals before transforming back.

ED50 is ‘effective dose for 50% success’ and LD50 is ‘50% lethal dose’.

# Poisson Regression

Suppose that rather than success/failure, we have the number of events that happened, for example the number of insurance claims. Then the natural model is that  $Y_i \sim \text{Poisson}(\mu_i)$ , or perhaps that  $Y_i \sim \text{Poisson}(\lambda_i T_i)$  where  $T_i$  is the time at risk.

We want the mean  $\mu_i$  or rate  $\lambda_i$  to depend on  $p$  carriers, and it is (for all the usual reasons) most natural to work on log scale:

$$\log \mu = \eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

or

$$\log \lambda = \eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

$$\log \mu = \eta + \log T = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \log T$$

The term  $\log T$  is known as an *offset*, a regression term with coefficient fixed at one.

## Poisson Regression 2

Just as for binomial logistic regression we can group observations with the same values of the regressors, write down a likelihood, maximize it numerically (and be careful about non-existence of MLEs) and use asymptotic theory (noting that the analogue of the Hauck–Donner effect holds) and differences in deviances as LRTs.

Once again there are several types of residuals and they do not work well if the counts are small.

The residual deviance can be used as a measure of fit, but note that the asymptotic  $\chi^2_{M-p}$  distribution is pretty inaccurate if the means are small. (See later.)

# Log-linear Models

In one sense *log-linear models* are just Poisson models with a log-linear model for the means (or rates).

However, they are usually referring to a contingency table, in which we have a  $K > 2$ -way response and a series of categorical explanatory variables. Then the data are a count  $Y_{ik}$  for response level  $k$  from the cell with the  $i$ th combination of explanatory variables. Under Poisson sampling,  $Y_{ik} \sim \text{Poisson}(\mu_{ik})$ , independently for each  $i$  and  $k$ . This is just a change of notation, and we can write

$$\log \mu_{ik} = \eta_{ik} = \beta_k + \alpha_i + \beta_{ik}$$

where we will expand out  $\beta_{ik}$  as interactions between the explanatory variables defining  $i$  and the response factor with levels  $k$ .

Note that we can take  $\beta_1 = 0$  and  $\beta_{i1} = 0$ . Similarly we can have  $\beta_{1k} = 0$ .

# Log-linear and Multiple Logistic Models

Now suppose we condition on the total number of observations for each cell  $i$  — this is product multinomial sampling. Then the joint distribution of  $(Y_{i1}, \dots, Y_{iK})$  is multinomial (independent for each  $i$ ) with probabilities

$$p_{ik} = \mu_{ik} / \sum_{\ell=1}^K \mu_{\ell} = e^{\eta_{ik}} / \sum_{\ell} e^{\eta_{i\ell}} = e^{\beta_k + \alpha_i + \beta_{ik}} / \sum_{\ell} e^{\beta_{\ell} + \alpha_i + \beta_{i\ell}} = e^{\beta_k + \beta_{ik}} / \sum_{\ell} e^{\beta_{\ell} + \beta_{i\ell}}$$

This is the generalization to  $K > 2$  responses of logistic regression, and is variously known as a log-linear, multiple logistic or multinomial logistic model.

This is a linear model for the log-odds of response  $k$  vs response 1, for

$$\log p_{ik}/p_{i1} = \beta_k + \beta_{ik} - \beta_1 + \beta_{i1}$$

and in particular

$$\log p_{ik}/p_{i1} = \beta_k + \beta_{ik}$$

# Inference for Product-Multinomial Models

The only differences from the Poisson case are that the likelihood is different (it is conditional) and the parameters  $\alpha_i$  are no longer estimable. However, the MLEs for  $\beta_{ik}$  occur at the same values, the observed and Fisher information for those parameters is the same, as is the change in deviance in adding or subtracting terms.

It used to be computationally convenient to assume Poisson sampling even when product multinomial sampling was true, although this does imply fitting  $(\alpha_i)$ , potentially a lot of extra parameters. This is known as *fitting a surrogate Poisson model*.

Note that in particular fitting a 2-level response by a product-multinomial log-linear model is equivalent to fitting by logistic regression.

# Example: Copenhagen Housing Satisfaction

We saw a four-way classification of 1 681 householders in Copenhagen who were surveyed on the *type* of rental accommodation they occupied, the degree of *contact* they had with other residents, their feeling of *influence* on apartment management and their level of *satisfaction* with their housing conditions.

We have a table of counts – one could consider influence and satisfaction to be ordered factors.

If we are only interested in satisfaction, we need to include terms in the model to account for all the *history* variables, so the *minimal model* is  $\text{Infl*Type*Cont}$ , and the interest is in interactions between Sat and the explanatory variables.

<b>Contact</b>		<b>Low</b>			<b>High</b>		
<b>Satisfaction</b>		<b>Low</b>	<b>Med.</b>	<b>High</b>	<b>Low</b>	<b>Med.</b>	<b>High</b>
<b>Housing</b>	<b>Influence</b>						
Tower blocks	Low	21	21	28	14	19	37
	Medium	34	22	36	17	23	40
	High	10	11	36	3	5	23
Apartments	Low	61	23	17	78	46	43
	Medium	43	35	40	48	45	86
	High	26	18	54	15	25	62
Atrium houses	Low	13	9	10	20	23	20
	Medium	8	8	12	10	22	24
	High	6	7	9	7	10	21
Terraced houses	Low	18	6	7	57	23	13
	Medium	15	13	13	31	21	13
	High	7	5	11	5	6	13

```
> names(housing)
[1] "Sat"  "Infl" "Type" "Cont" "Freq"
> house.glm0 <- glm(Freq ~ Infl*Type*Cont + Sat,
                    family = poisson, data = housing)
> summary(house.glm0, cor = F)
. . . .
Null Deviance: 833.66 on 71 degrees of freedom
Residual Deviance: 217.46 on 46 degrees of freedom
```

The high residual deviance clearly indicates that this simple model is inadequate, so the probabilities do appear to vary with the explanatory factors.

We now consider adding the simplest terms to the model.

```
> addterm(house.glm0, ~. + Sat:(Infl+Type+Cont), test = "Chisq")
....
      Df Deviance    AIC    LRT Pr(Chi)
<none>      217.46 269.46
Sat:Infl   4   111.08 171.08 106.37 0.00000
Sat:Type   6   156.79 220.79  60.67 0.00000
Sat:Cont   2   212.33 268.33   5.13 0.07708
```

It turns out that all three terms are necessary, so we now update our initial model by including all three at once.

```
> house.glm1 <- update(house.glm0, . ~ . + Sat:(Infl+Type+Cont))
> summary(house.glm1, cor = F)
....
Null Deviance: 833.66 on 71 degrees of freedom
Residual Deviance: 38.662 on 34 degrees of freedom
```

The deviance indicates a satisfactorily fitting model, but we might look to see if some adjustments to the model might be warranted.

```
> dropterm(house.glm1, test = "Chisq")
```

```
.....
```

	Df	Deviance	AIC	LRT	Pr(Chi)
<none>		38.66	114.66		
Sat:Infl	4	147.78	215.78	109.12	0.00000
Sat:Type	6	100.89	164.89	62.23	0.00000
Sat:Cont	2	54.72	126.72	16.06	0.00033
Infl:Type:Cont	6	43.95	107.95	5.29	0.50725

Note that the final term here is part of the minimum model and hence may *not* be removed. Only terms that contain the response factor, Sat, are of any interest to us for this analysis.

Now consider adding possible interaction terms.

```
> addterm(house.glm1, ~. + Sat:(Infl+Type+Cont)^2, test = "Chisq")
```

```
.....
```

	Df	Deviance	AIC	LRT	Pr(Chi)
<none>		38.662	114.66		
Sat:Infl:Type	12	16.107	116.11	22.555	0.03175
Sat:Infl:Cont	4	37.472	121.47	1.190	0.87973
Sat:Type:Cont	6	28.256	116.26	10.406	0.10855

The first term, a type  $\times$  influence interaction, appears to be mildly significant, but as it increases the AIC we choose not include it on the grounds of simplicity. We have now shown (subject to checking assumptions) that the three explanatory factors, type, influence and contact do affect the probabilities of each of the satisfaction classes in a simple, ‘main effect’-only way.

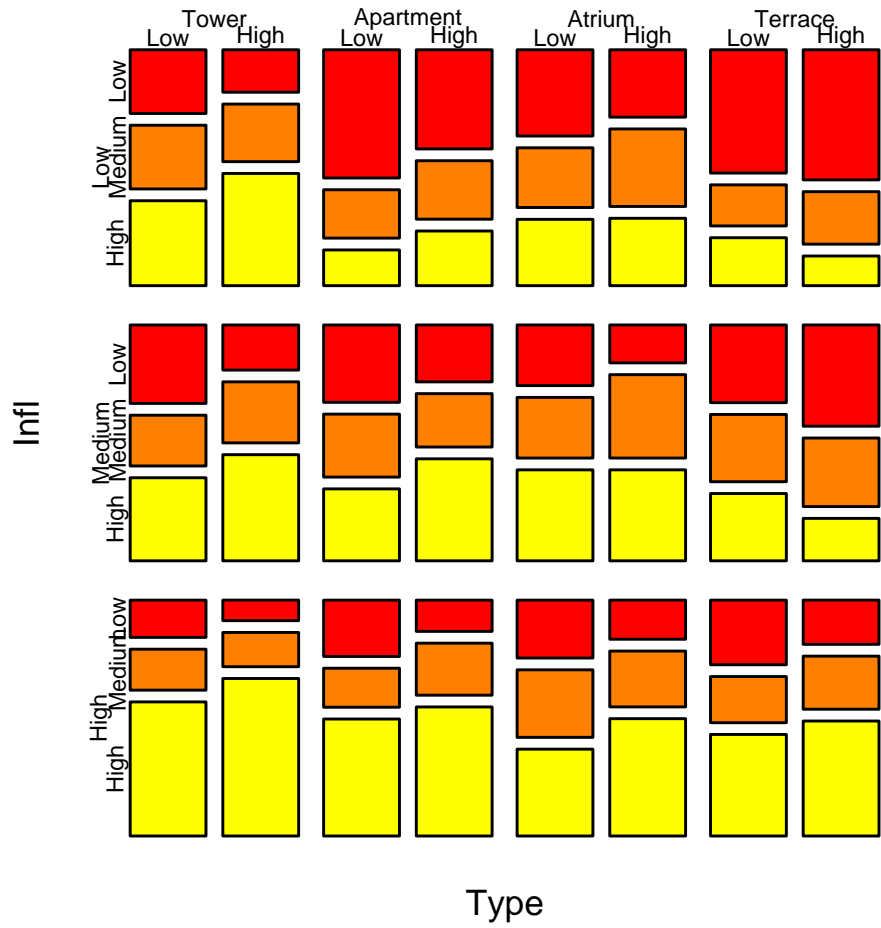
Our next step is to look at these estimated probabilities under the model and assess what effect these factors are having. The picture becomes clear if we normalize the means to probabilities.

```
hnames <- lapply(housing[, -5], levels) # omit Freq
house.pm <- predict(house.glm1, expand.grid(hnames),
                   type = "response") # poisson means
house.pm <- matrix(house.pm, ncol = 3, byrow = T,
                  dimnames = list(NULL, hnames[[1]]))
house.pr <- house.pm/drop(house.pm %*% rep(1, 3))
cbind(expand.grid(hnames[-1]), prob = round(house.pr, 2))
```

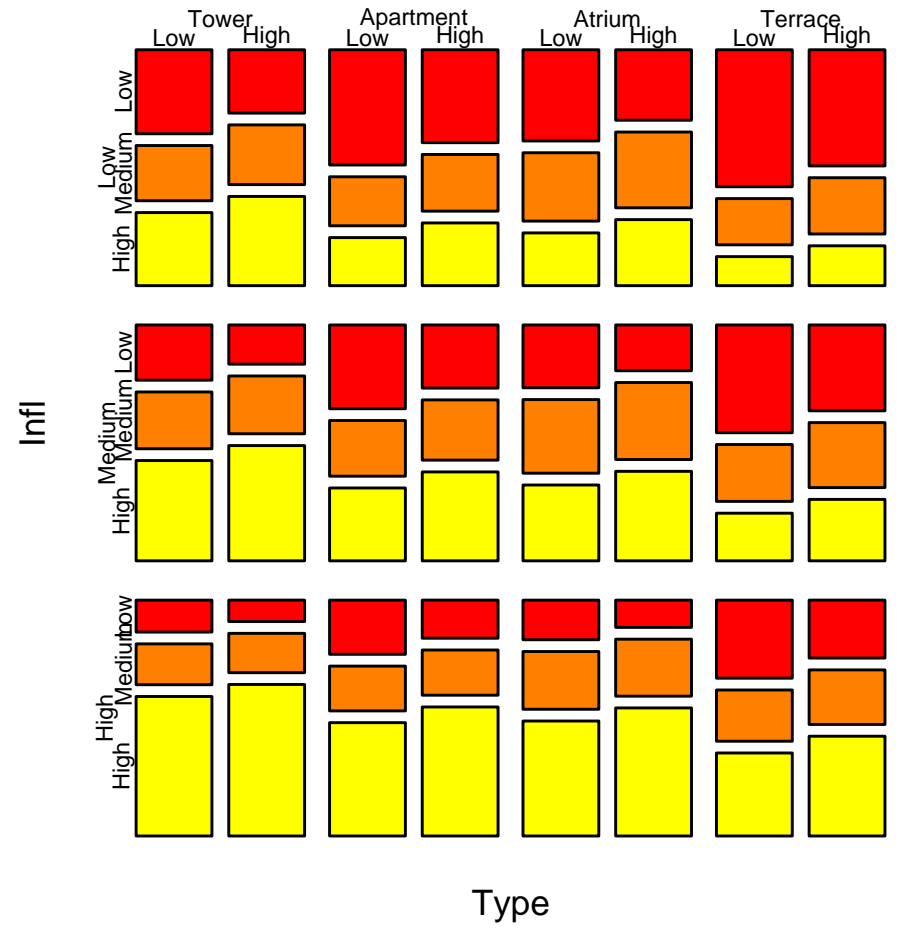
We might want to show this as a mosaic plot.

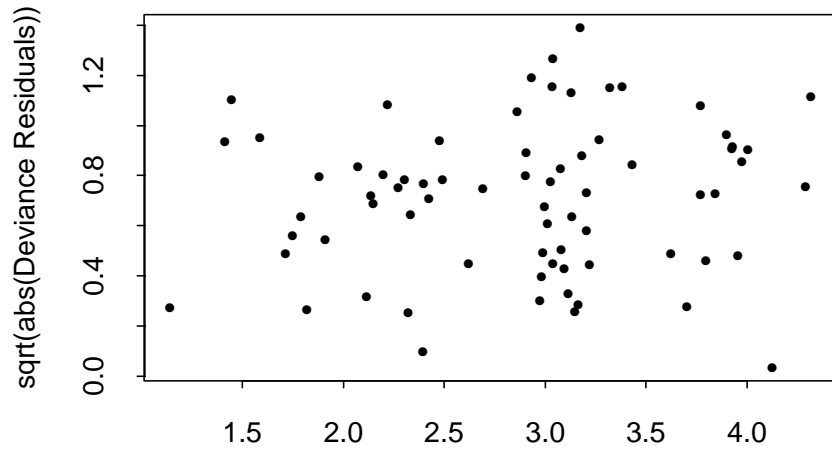
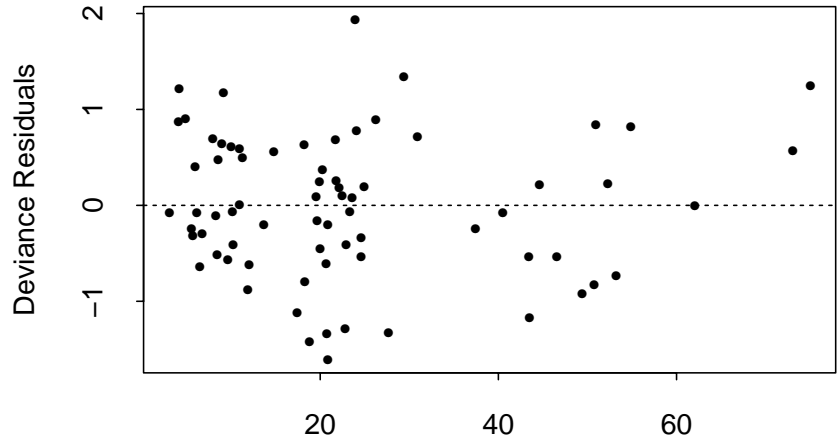
<b>Contact</b>		<b>Low</b>			<b>High</b>		
		<b>Low</b>	<b>Med.</b>	<b>High</b>	<b>Low</b>	<b>Med.</b>	<b>High</b>
<b>Housing</b>	<b>Influence</b>						
Tower blocks	Low	0.40	0.26	0.34	0.30	0.28	0.42
	Medium	0.26	0.27	0.47	0.18	0.27	0.54
	High	0.15	0.19	0.66	0.10	0.19	0.71
Apartments	Low	0.54	0.23	0.23	0.44	0.27	0.30
	Medium	0.39	0.26	0.34	0.30	0.28	0.42
	High	0.26	0.21	0.53	0.18	0.21	0.61
Atrium houses	Low	0.43	0.32	0.25	0.33	0.36	0.31
	Medium	0.30	0.35	0.36	0.22	0.36	0.42
	High	0.19	0.27	0.54	0.13	0.27	0.60
Terraced houses	Low	0.65	0.22	0.14	0.55	0.27	0.19
	Medium	0.51	0.27	0.22	0.40	0.31	0.29
	High	0.37	0.24	0.39	0.27	0.26	0.47

### Empirical probs

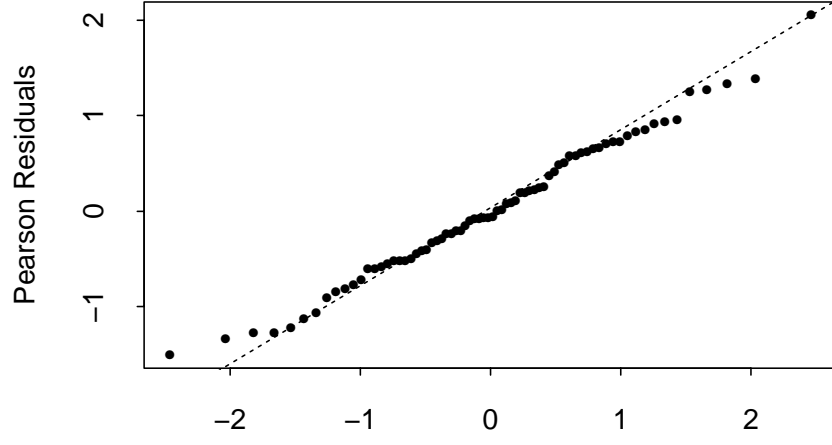
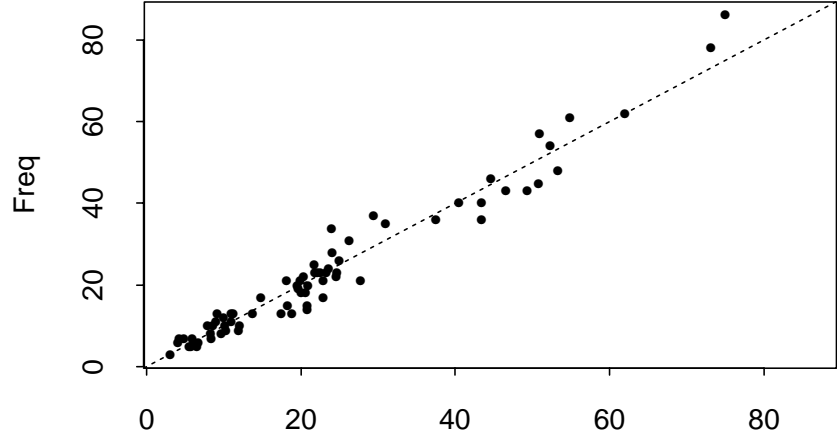


### Fitted probs





ted : Infl + Type + Cont + Sat + Infl:Type + Infl:Cont + Type:Cont + Sat:lr



ted : Infl + Type + Cont + Sat + Infl:Type + Infl:Cont + Type:Cont + Sat:lr

Quantiles of Standard Normal

# Fitting as a Multinomial Logistic Model

We can fit a multinomial model directly rather than use a surrogate Poisson model by using our function. No interactions are needed:

```
> library(nnet)
> house.mult <- multinom(Sat ~ Infl + Type + Cont,
                          weights = Freq, data = housing)
> house.mult
Coefficients:
      (Intercept) InflMedium InflHigh TypeApartment TypeAtrium
Medium    -0.41923    0.44644    0.66497    -0.43564     0.13134
High      -0.13871    0.73488    1.61264    -0.73566    -0.40794
      TypeTerrace      Cont
Medium    -0.66658  0.36085
High      -1.41234  0.48188

Residual Deviance: 3470.10    AIC: 3498.10
```

Here the deviance is comparing with the model that correctly predicts each person, not the multinomial response for each cell of the minimum model: we can compare with the usual saturated model by

```
> house.mult2 <- multinom(Sat ~ Infl*Type*Cont,  
                           weights = Freq, data = housing)  
> anova(house.mult, house.mult2, test = "none")
```

....

	Model	Resid. df	Resid. Dev	Test	Df	LR stat.
1	Infl + Type + Cont	130	3470.1			
2	Infl * Type * Cont	96	3431.4	1 vs 2	34	38.662

# Logistic Models for Ordinal Responses

Now suppose the  $K > 2$  responses are ordered. Potentially we can use this to simplify the model. For each  $k = 1, \dots, K - 1$  we can dichotomize the response into  $Y > k$  versus  $Y \leq k$ . Suppose that the log-odds ratios

$$\text{logit}P(Y > k \mid \text{cell } i) = -\gamma_k + \beta_i$$

differ only by a constant for different  $k$  across all cells  $i$ . Then we have a *proportional odds logistic models* and this has  $K - 1$  intercepts but only one (rather than  $K - 1$ ) sets of  $\beta$ 's. (Note that the  $\gamma_k$ 's must increase with  $k$ , hence the negative sign.)

A POLR is a more restrictive model with many fewer parameters than the full multiple logistic regression model. Once again it is fitted by numerically maximizing the likelihood, changes in deviance give LRTs, and so on.

There is another instructive way to look at POLRs. Suppose  $Z_i$  is an underlying continuous response for cell  $i$ , and that we actually observe is a ‘*coarsened*’ version of it, so

$$Y_i = k \quad \text{if and only if} \quad \zeta_{k-1} < Z_i < \zeta_k$$

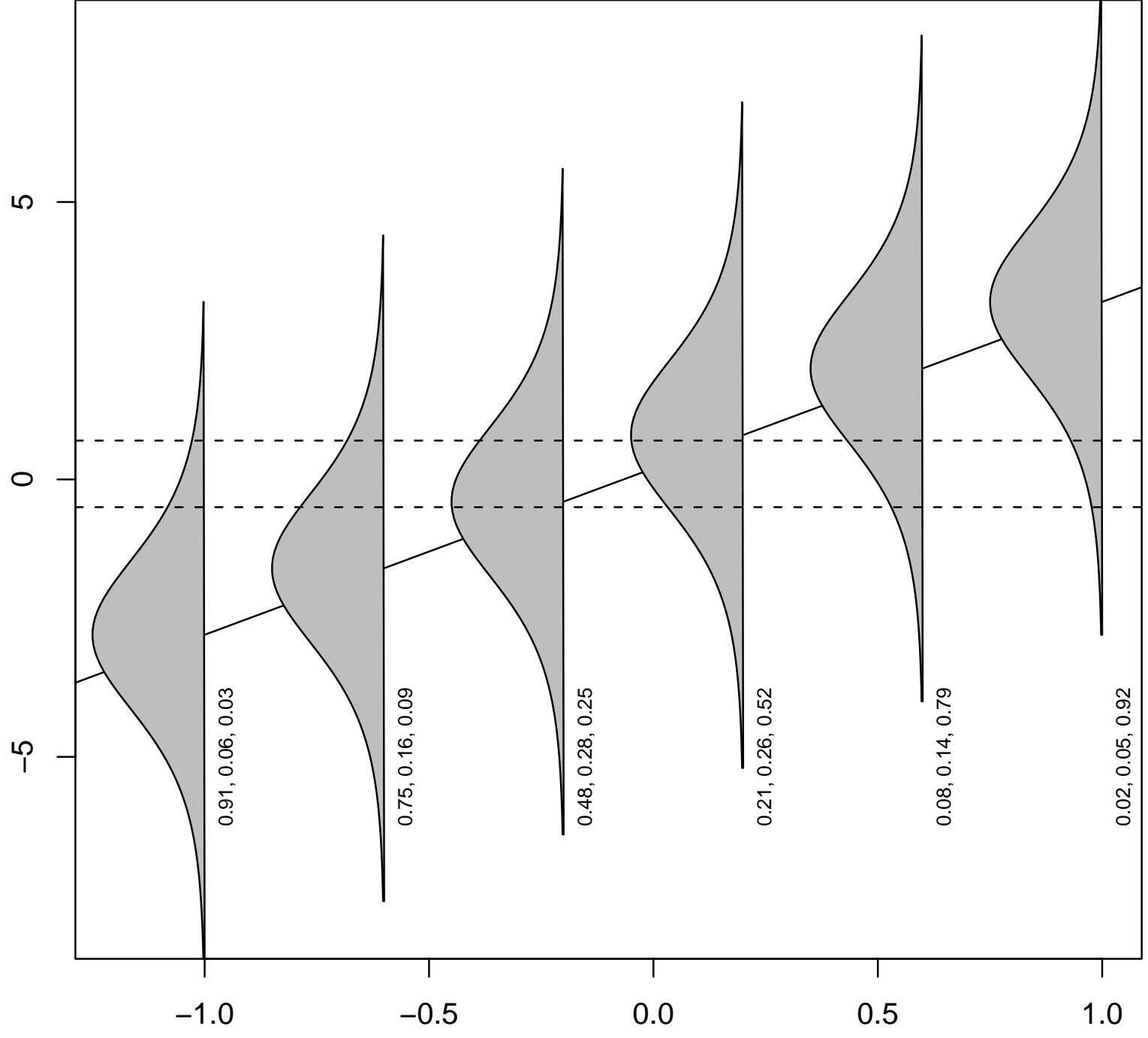
Now suppose  $Z_i$  has a logistic distribution with mean  $\beta_i$ .

$$P(Y_i > \zeta_k) = 1 - \frac{e^{\zeta_k - \beta_i}}{1 + e^{\zeta_k - \beta_i}} = \frac{1}{1 + e^{\zeta_k - \beta_i}} = \frac{e^{-\zeta_k + \beta_i}}{1 + e^{-\zeta_k + \beta_i}}$$

so we have a response greater than  $k$  if and only if  $P(Y_i > \zeta_k)$  and

$$\text{logit}P(Y > k \mid \text{cell } i) = -\zeta_k + \beta_i$$

Note that the actual distribution of  $Y$  does not matter as we can use any non-linear monotonic transformation of it, but the linear model for the means only hold for the logistic scale.



## Fitting a POLR to the Housing Data

We can look at the two possible cumulative logit models and compare the cumulative logits from our fitted probabilities by

```
> house.cpr <- apply(house.pr, 1, cumsum)
> logit <- function(x) log(x/(1-x))
> house.ld <- logit(house.cpr[2, ]) - logit(house.cpr[1, ])
> sort(drop(house.ld))
 [1] 0.93573 0.98544 1.05732 1.06805 1.07726 1.08036 1.08249
 [8] 1.09988 1.12000 1.15542 1.17681 1.18664 1.20915 1.24350
[15] 1.27241 1.27502 1.28499 1.30626 1.31240 1.39047 1.45401
[22] 1.49478 1.49676 1.60688
> mean(.Last.value)
 [1] 1.2238
```

The average log-odds ratio is about 1.2 and variations from it are not great. So a POLR model looks plausible.

```

> (house.plr <- polr(Sat ~ Infl + Type + Cont,
                    data = housing, weights = Freq))
    ....
Coefficients:
  InflMedium InflHigh TypeApartment TypeAtrium TypeTerrace
    0.56639    1.2888      -0.57234    -0.36619      -1.091
  Cont
  0.36029

```

```

Intercepts:
  Low|Medium Medium|High
   -0.49612    0.69072

```

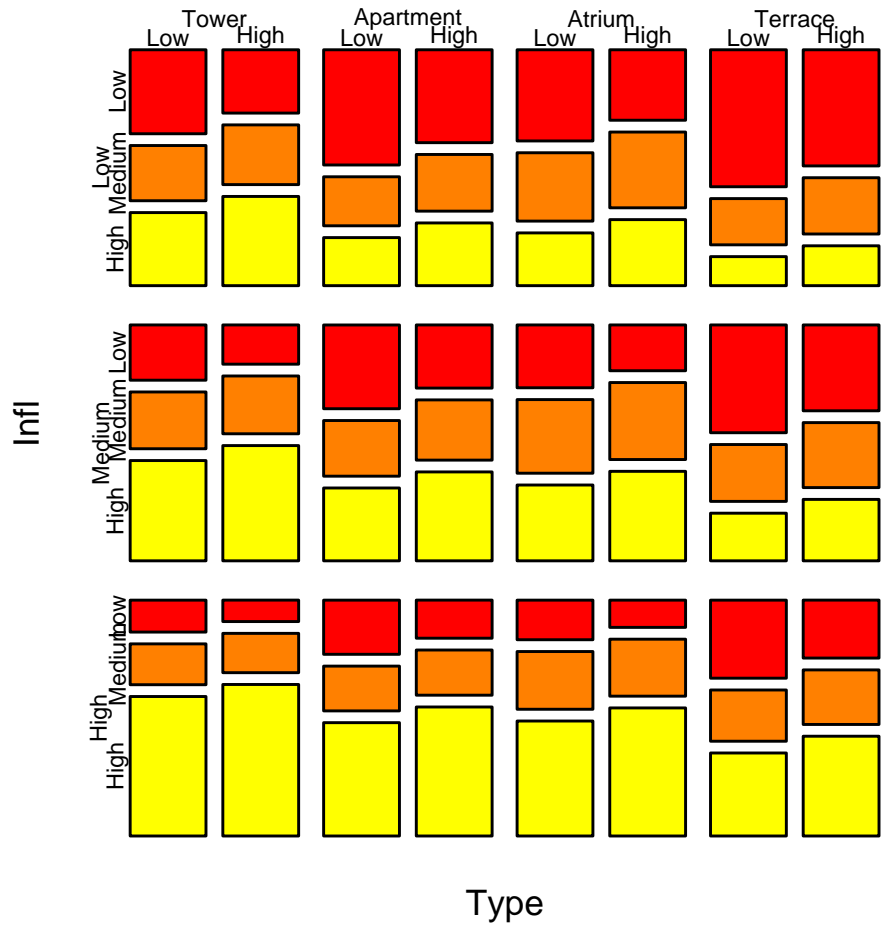
```

Residual Deviance: 3479.10    AIC: 3495.10

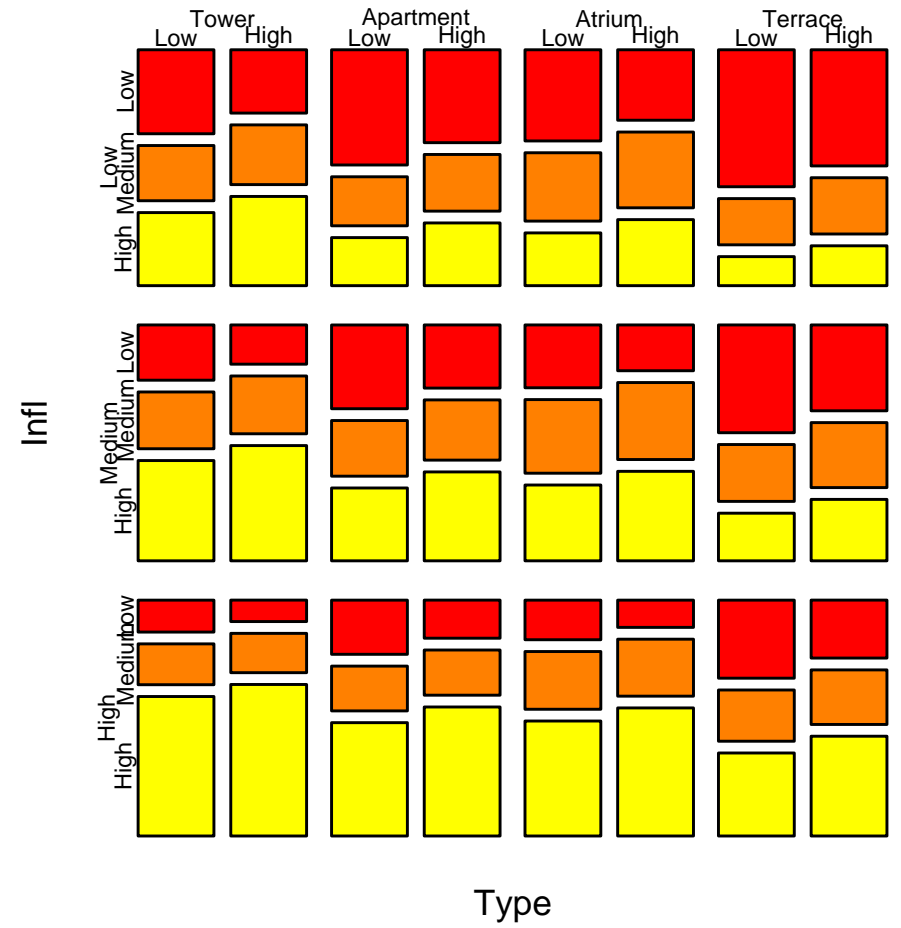
```

The residual deviance is comparable with that of the multinomial model fitted before, showing that the increase is only 9.0 for a reduction from 14 to 8 parameters. The AIC criterion is consequently much reduced.

### Multinomial logistic



### Proportional odds logistic



Note that the difference in intercepts, 1.19, agrees fairly well with the average logit difference of 1.22 found previously.

The fitted probabilities are close to those of the multinomial model.

The advantage of the proportional-odds model is not just that it is so much more parsimonious than the multinomial, but with the smaller number of parameters the action of the covariates on the probabilities is much easier to interpret and to describe. However, as it is more parsimonious, stepAIC will select a more complex model linear predictor:

```
> house.plr2 <- stepAIC(house.plr, ~.^2)
> house.plr2$anova
```

```
.....
      Step Df Deviance Resid. Df Resid. Dev    AIC
1                1673      3479.1 3495.1
2 + Infl:Type    6    22.509    1667      3456.6 3484.6
3 + Type:Cont    3     7.945    1664      3448.7 3482.7
```

# Introduction to GLMs

The similarity between binomial logistic and Poisson log-linear models is not coincidental: they are both members of the class of *generalized linear models*, although there are few other useful examples.

- There is a response  $y$  observed independently at fixed values of stimulus variables  $x_1, \dots, x_p$ .
- The stimulus variables may only influence the distribution of  $y$  through a single linear function called the *linear predictor*  $\eta = \beta_1 x_1 + \dots + \beta_p x_p$
- The distribution of  $y$  has density of the form

$$f(y_i; \theta_i, \varphi) = \exp [A_i \{y_i \theta_i - \gamma(\theta_i)\} / \varphi + \tau(y_i, \varphi / A_i)]$$

where  $\varphi$  is a *scale parameter* (possibly known),  $A_i$  is a *known* prior weight and parameter  $\theta_i$  depends upon the linear predictor.

- The mean  $\mu$  is a smooth invertible function of the linear predictor:

$$\mu = m(\eta), \quad \eta = m^{-1}(\mu) = \ell(\mu)$$

The inverse function  $\ell(\cdot)$  is called the *link function*.

Note that  $\theta$  is also an invertible function of  $\mu$ , and the variance is a function of the mean (and hence of  $\theta$ ). If  $\varphi$  were known the distribution of  $y$  would be a one-parameter canonical exponential family. An unknown  $\varphi$  is handled as a nuisance parameter by moment methods.

GLMs allow a unified treatment of statistical methodology for several important classes of models. We consider a few examples.

**Gaussian** For a normal distribution  $\varphi = \sigma^2$  and we can write

$$\log f(y) = \frac{1}{\varphi} \left\{ y\mu - \frac{1}{2}\mu^2 - \frac{1}{2}y^2 \right\} - \frac{1}{2} \log(2\pi\varphi)$$

so  $\theta = \mu$  and  $\gamma(\theta) = \theta^2/2$ .

**Poisson** For a Poisson distribution with mean  $\mu$  we have

$$\log f(y) = y \log \mu - \mu - \log(y!)$$

so  $\theta = \log \mu$ ,  $\varphi = 1$  and  $\gamma(\theta) = \mu = e^\theta$ .

**Binomial** For a binomial distribution with fixed number of trials  $a$  and parameter  $p$  we take the response to be  $y = s/a$  where  $s$  is the number of ‘successes’. The density is

$$\log f(y) = a \left[ y \log \frac{p}{1-p} + \log(1-p) \right] + \log \binom{a}{ay}$$

so we take  $A_i = a_i$ ,  $\varphi = 1$ ,  $\theta$  to be the logit transform of  $p$  and  $\gamma(\theta) = -\log(1-p) = \log(1+e^\theta)$ .

Several links have been proposed, but for each distribution the *canonical link* gives a canonical exponential family.

Family	Canonical link	Name	Variance	Name
binomial	$\log(\mu/(1 - \mu))$	logit	$\mu(1 - \mu)$	mu(1-mu)
Gamma	$-1/\mu$	inverse	$\mu^2$	mu^2
gaussian	$\mu$	identity	1	constant
inverse.gaussian	$-2/\mu^2$	1/mu^2	$\mu^3$	mu^3
poisson	$\log \mu$	log	$\mu$	mu

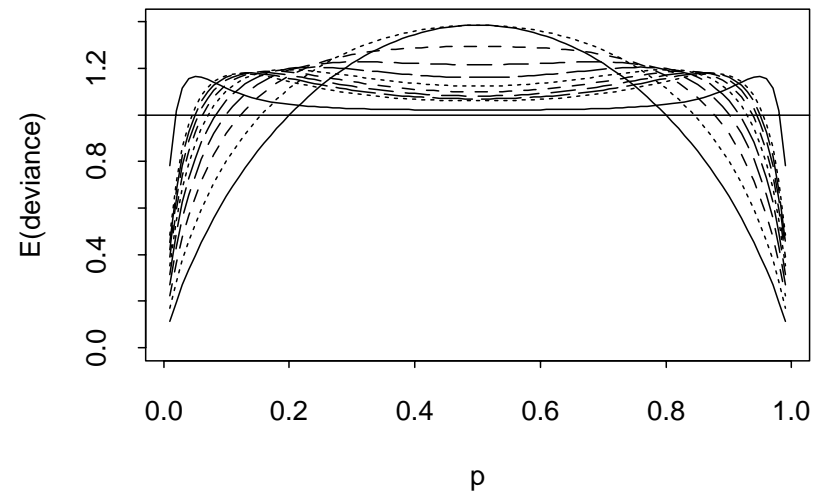
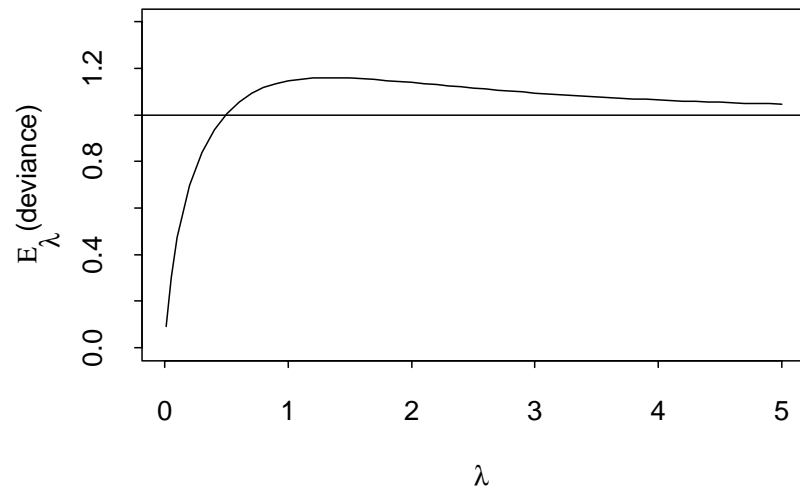
# Over-dispersion

The role of the dispersion parameter  $\varphi$  in the theory and practice of GLMs is often misunderstood.

For binomial and Poisson families the theory specifies that  $\varphi = 1$ , but in some cases we estimate  $\varphi$  as if it were an unknown parameter and use that value in standard error calculations and as a denominator in approximate  $F$ -tests rather than use chi-squared tests. This is an *ad hoc* adjustment for over-dispersion but the corresponding likelihood may not correspond to any family of error distributions.

However, if over-dispersion is present and not adjusted for, severely over-optimistic inferences may result.

A common way to ‘discover’ over- or under-dispersion is to notice that the residual deviance is appreciably different from the residual degrees of freedom, since in the usual theory the expected value of the residual deviance should equal the degrees of freedom. The theory is asymptotic, and only applies for large  $n_i p_i$  for a binomial and for large  $\mu_i$  for a Poisson.



Plots of the expected residual deviance against (left) the parameter of a Poisson and (right) the  $p$  for a binomial( $n, p$ ) for  $n = 1, 2, \dots, 10, 25$ .