University of Glasgow

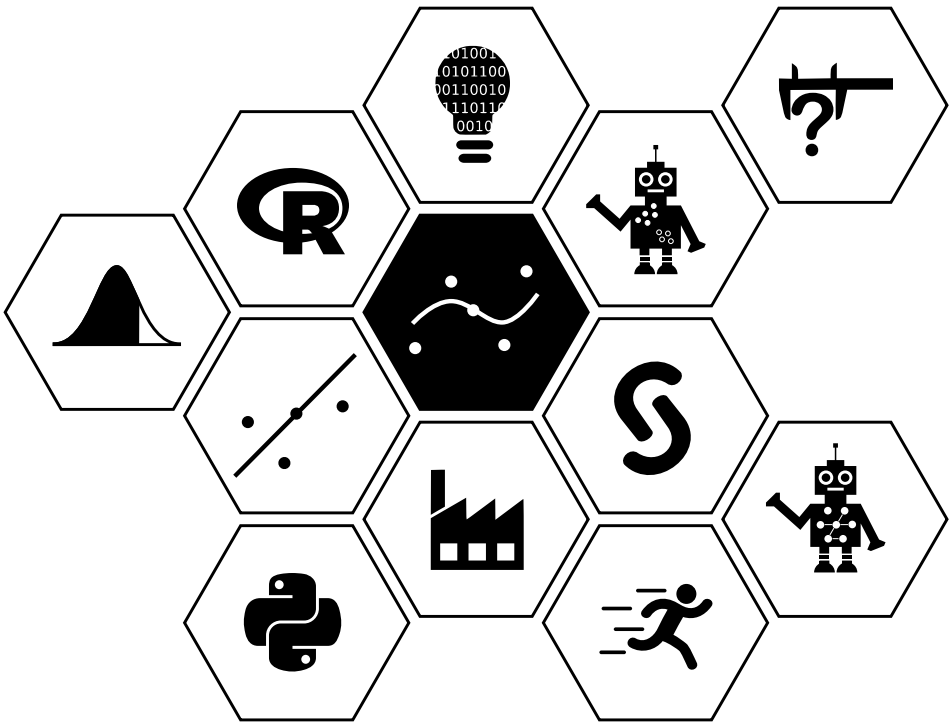
Advanced Predictive Models

Tereza Neocleous

Academic Year 2020-21

Week 4:

Models for nominal/ordinal response



DATA ANALYTICS GLASGOW

Models for categorical responses

In last week's material we covered GLMs for categorical responses with two possible outcomes. This week, we will generalise this to situations where the response variable is categorical with more than two categories. More specifically, we will look at logistic regression models applied to **nominal** (unordered) or **ordinal** (ordered) responses with **more than two categories**.

The basis for modelling categorical data with more than two categories is the **multinomial distribution**.

Multinomial distribution

Definition 1.

*Consider a random variable Y with J categories. Let $p_1, p_2, \ldots, p_J$ be the respective probabilities associated with each of the J categories, with $p_1 + p_2 + \cdots + p_J = 1$. Suppose there are n independent observations which result in y_1 outcomes in category 1, y_2 outcomes in category 2, and so on. Let $\mathbf{y} = (y_1, y_2, \ldots, y_J)^\intercal$ with $\sum_{j=1}^{J} y_j = n$. We say that $\mathbf{y}$ follows a **multinomial distribution** with probability mass function (p.m.f.)*

$$f(\mathbf{y}|n) = \frac{n!}{y_1! y_2! \ldots y_J!} p_1^{y_1} p_2^{y_2} \ldots p_J^{y_J}. \tag{1}$$

Properties of the multinomial distribution

If $J = 2$ then $p_2 = 1 - p_1$ and $y_2 = n - y_1$ so the expression above reduces to the p.m.f. of the binomial distribution:

$$f(\mathbf{y}|n) = \frac{n!}{y_1!(n - y_1)!} p_1^{y_1} p_2^{n-y_1}$$

For the multinomial distribution, we have the following expressions for the mean, variance and covariance:

$$\begin{aligned} \mathrm{E}(Y_j) &= np_j \\ \mathrm{Var}(Y_j) &= np_j(1 - p_j) \\ \mathrm{Cov}(Y_j, Y_k) &= -np_j p_k \end{aligned}$$

Notice that for $J = 2$ you obtain the mean and variance for a binomial random variable. Notice also the negative covariance between Y_j and Y_k due to the sum constraint $\sum_{j=1}^{J} y_j = n$.

In general, equation (1) does not satisfy the exponential family distribution requirement for the response in a GLM, but we can still fit GLMs to multinomial responses thanks to the following relationship with the Poisson distribution, which is a member of the exponential family.

Relationship with Poisson distribution

The multinomial distribution is not a member of the exponential family. However, we can still use the multinomial distribution in the GLM context if we view it as the joint distribution of Poisson random variables conditional on their sum n.

Supplementary material:

Let $Y_j \sim \text{Po}(\mu_j)$ where the Y_j are independent for $j = 1, \ldots, J$. Their joint p.m.f. is:

$$f(\mathbf{y}) = \prod_{j=1}^{J} \frac{\mu_j^{y_j} e^{-\mu_j}}{y_j!}$$

The random variable $n = Y_1 + Y_2 + \cdots + Y_J$ follows the $\text{Po}(\mu_1 + \mu_2 + \cdots + \mu_J)$ distribution.

Conditional on n, $\mathbf{y}$ has the following distribution:

$$f(\mathbf{y}|n) = \frac{\prod_{j=1}^{J} \mu_j^{y_j} e^{-\mu_j} / y_j!}{(\mu_1 + \mu_2 + \cdots + \mu_J)^n e^{-(\mu_1+\mu_2+\cdots+\mu_J)} / n!}$$

This can be simplified to:

$$f(\mathbf{y}|n) = \frac{n!}{y_1! y_2! \ldots y_J!} \left(\frac{\mu_1}{\sum \mu_k}\right)^{y_1} \left(\frac{\mu_2}{\sum \mu_k}\right)^{y_2} \ldots \left(\frac{\mu_J}{\sum \mu_k}\right)^{y_J}$$

which is the same as the expression for the multinomial p.m.f. in equation (1), if we let $p_j = \dfrac{\mu_j}{\sum_{k=1}^{J} \mu_k}$.

Nominal logistic regression

Nominal logistic regression, also known as *multinomial logistic regression* is used when there is no natural order among the response categories, for example:

- Eye colour: Blue, Green, Brown, Hazel
- House types: Bungalow, Duplex, Terrace
- Type of pet: Dog, Cat, Rodent, Fish, Bird
- Genotype: AA, Aa, aa

One category is arbitrarily chosen as the reference category, and all other categories are compared with it. Suppose the first category is chosen as the reference category. Then the logits for the other categories are defined by

$$\text{logit}(p_j) = \log\left(\frac{p_j}{p_1}\right) = \mathbf{x}^\top \boldsymbol{\beta}_j, \qquad \text{for } j = 2, \ldots, J. \tag{2}$$

Parameter estimation and fitted values

The $J - 1$ logit equations (2) are solved simultaneously to estimate the parameters $\boldsymbol{\beta}_j$.

Given parameter estimates $\hat{\boldsymbol{\beta}}_j$, the linear predictors $\mathbf{x}^\top \hat{\boldsymbol{\beta}}_j$ can be calculated. From equation (2), $\hat{p}_j = \hat{p}_1 \exp(\mathbf{x}^\top \hat{\boldsymbol{\beta}}_j)$.

Since $\hat{p}_1 + \hat{p}_2 + \cdots + \hat{p}_J = 1$,

$$\hat{p}_1 = \frac{1}{1 + \sum_{j=2}^{J} \exp(\mathbf{x}^\top \hat{\boldsymbol{\beta}}_j)} \tag{3}$$

and

$$\hat{p}_j = \frac{\exp(\mathbf{x}^\top \hat{\boldsymbol{\beta}}_j)}{1 + \sum_{j=2}^{J} \exp(\mathbf{x}^\top \hat{\boldsymbol{\beta}}_j)}. \tag{4}$$

Fitted values (expected frequencies) can be calculated for each covariate pattern by multiplying the estimated probabilities $\hat{p}_j$ by the total frequency of the covariate pattern.

Parameter estimates $\hat{\boldsymbol{\beta}}_j$ depend on the choice of reference category, but fitted values don't.

Model checking and model comparisons

Summary statistics can be used to assess the adequacy of a model and also to compare models. Some of the statistics we can consider are:

- the **deviance** $D = 2[l(\hat{\boldsymbol{\beta}}_{\max}) - l(\hat{\boldsymbol{\beta}})]$ (also referred to as *residual deviance*), where $l(\hat{\boldsymbol{\beta}}_{\max})$ is the maximised log-likelihood for the saturated (full) model and $l(\hat{\boldsymbol{\beta}})$ is the maximised log-likelihood for the model of interest;
- the **likelihood ratio statistic**, which is equal to the difference between the residual deviance for the model of interest and the null deviance (deviance of the model with no predictors included);
- the **Akaike information criterion** $AIC = -2l(\hat{\boldsymbol{\beta}}; \mathbf{y}) + 2p$, which equals the maximised log-likelihood of the model of interest plus a penalty term equal to twice the number of parameters in the model. The reason for this is that we can keep adding predictors to the model to improve the log-likelihood, but the cost is increased model complexity. The penalty term attempts to strike a balance between model complexity and how well the model fits.

If the model fits well, the deviance will be asymptotically $\chi^2(N - p)$, where N is $J - 1$ times the number of distinct covariate patterns in the data, and p is the number of parameters estimated.

The likelihood ratio statistic will be asymptotically $\chi^2[p - (J - 1)]$ because the null (minimal) model will have one parameter for each logit defined in equation (2).

The AIC can be used for model selection: calculate the criterion for each model and choose the one with the smallest value of the AIC.

Coefficient interpretation in terms of odds ratios

Consider a response with J categories and a single explanatory variable x which denotes whether an exposure factor is present ($x = 1$) or not ($x = 0$). Let $p_{j,\text{present}}$ be the probability of the jth category assuming the exposure is present, $p_{j,\text{absent}}$ the probability of the jth category assuming the exposure is absent. The odds ratio for exposure j, where $j = 2, \ldots, J$, relative to the reference category $j = 1$ is

$$OR_j = \frac{p_{j,\text{present}}/p_{j,\text{absent}}}{p_{1,\text{present}}/p_{1,\text{absent}}}$$

The model

$$\log\left(\frac{p_j}{p_1}\right) = \beta_{0j} + \beta_{1j}x, \qquad j = 2, \ldots, J$$

gives log odds

$$\log\left(\frac{p_{j,\text{absent}}}{p_{1,\text{absent}}}\right) = \beta_{0j}, \qquad \text{when } x = 0$$
$$\log\left(\frac{p_{j,\text{present}}}{p_{1,\text{present}}}\right) = \beta_{0j} + \beta_{1j}, \qquad \text{when } x = 1$$

The log of the odds ratio can be written as

$$\log OR_j = \log\left(\frac{p_{j,\text{present}}}{p_{1,\text{present}}}\right) - \log\left(\frac{p_{j,\text{absent}}}{p_{1,\text{absent}}}\right) = \beta_{1j}$$

Hence, $OR_j = \exp(\beta_{1j})$ which we estimate by $\exp(\hat{\beta}_{1j})$. If $\beta_{1j} = 0$ (or equivalently $\exp(\beta_{1j}) = 1$), the exposure factor has no effect.

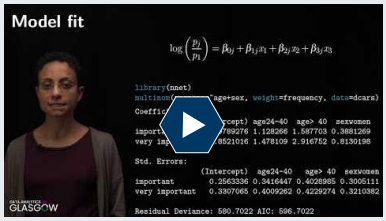
We can obtain 95% confidence intervals for OR_j using the formula:

$$\exp[\hat{\beta}_{1j} \pm 1.96\text{se}(\hat{\beta}_{1j})],$$

where 1.96 is the 97.5th percentile of the standard normal distribution.

Recall that this normal approximation is based on the asymptotic distribution of the MLE $\hat{\boldsymbol{\beta}}$ in a GLM, which is $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \mathcal{I}^{-1})$ where $\boldsymbol{\beta}$ is the true parameter vector and $\mathcal{I}^{-1}$ is the inverse of the information matrix.

A confidence interval which does not include 1 corresponds to a significant β.



Nominal logistic regression

https://youtu.be/SpQM2wTqpx8

Duration: 11m52s

Example 1 (Car preference data).

In this example we look at data on subjects that were interviewed about the importance of various features when buying a car.[a]

We focus in particular on the importance of power steering and air conditioning. The variables available in this dataset are:

- `sex`: woman/man
- `age`: 18-23, 24-40, >40
- `response`: no/little, important, very important

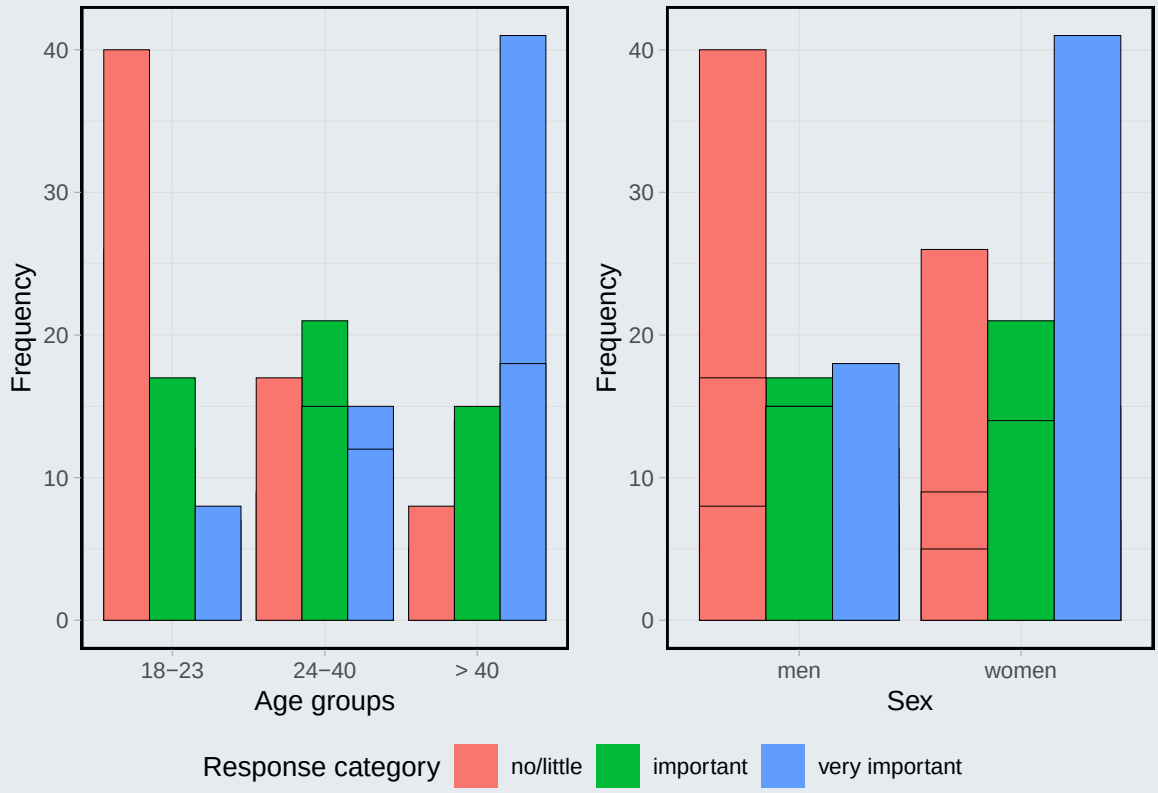
```
dcars <- read.csv(url("http://www.stats.gla.ac.uk/~tereza/rp/cars.csv"))
dcars$response <- factor(dcars$response,
                         levels = c("no/little", "important", "very important"))
dcars$age <- factor(dcars$age,
                         levels = c("18-23", "24-40", "> 40"))
```

	sex	age	response	frequency
1	women	18-23	no/little	26
2	women	18-23	important	12
3	women	18-23	very important	7
4	women	24-40	no/little	9
5	women	24-40	important	21
6	women	24-40	very important	15
7	women	> 40	no/little	5
8	women	> 40	important	14
9	women	> 40	very important	41
10	men	18-23	no/little	40
11	men	18-23	important	17
12	men	18-23	very important	8
13	men	24-40	no/little	17
14	men	24-40	important	15
15	men	24-40	very important	12
16	men	> 40	no/little	8
17	men	> 40	important	15
18	men	> 40	very important	18

From the plots of the data below, we can see that quite a large proportion of people – a little over 58% in the over 40 category considered the features *very important* and, similarly 60% of young people (18-23 years old) considered these features as having *no or little importance*. Sex also seems to have an impact on car feature preferences, with over 40% of men considering the features of *no or little importance* and over 40% of women considering them *very important*.

```
p1 <- ggplot(dcars, aes(x = age, y = frequency, fill = response)) +
      geom_bar(stat = "identity", position = "dodge" )+
      xlab("Age groups" ) + ylab("Frequency" ) +
      theme(legend.position = "none")
p2 <- ggplot(dcars, aes(x = sex, y = frequency, fill = response)) +
      geom_bar(stat = "identity", position = "dodge" )+
      xlab("Sex" ) + ylab("Frequency" ) +
```

```
  scale_fill_discrete(name = "Response category") +
  theme(legend.position = "bottom")
```



Although the response is really an ordinal variable, we will begin by treating it as nominal with "no/little importance" as the reference category (also occasionally referred to as "unimportant" in the rest for brevity.) Later on we will also fit an ordinal model. Similarly we will initially regard age as nominal.

We can fit the following **nominal logistic regression model** using the `multinom()` function from `library(nnet)`:

$$\log\left(\frac{p_j}{p_1}\right) = \beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2 + \beta_{3j}x_3, \qquad j = 2, 3,$$

where

- $j = 1$ for "no/little importance" (the reference category)
- $j = 2$ for "important"
- $j = 3$ for "very important"
- $x_1 = 1$ for women and 0 for men,
- $x_2 = 1$ for age 24-40 years and 0 otherwise
- $x_3 = 1$ for age $>$40 years and 0 otherwise.

```
m1 <- multinom(response ~ age + sex, weight = frequency, data = dcars)
```

```
# weights:  15 (8 variable)
initial  value 329.583687
iter  10 value 290.566455
final  value 290.351098
converged
```

```
summary(m1)
```

```
Call:
multinom(formula = response ~ age + sex, data = dcars, weights = frequency)

Coefficients:
               (Intercept) age24-40  age> 40  sexwomen
important       -0.9789276 1.128266 1.587703 0.3881269
very important  -1.8521016 1.478109 2.916752 0.8130198
```

```
Std. Errors:
               (Intercept)  age24-40   age> 40  sexwomen
important        0.2563336 0.3416447 0.4028985 0.3005111
very important   0.3307065 0.4009262 0.4229274 0.3210382

Residual Deviance: 580.7022
AIC: 596.7022
```

Notice the two sets of coefficients, for the categories "important" and "very important" that correspond to the two logit equations comparing these to the baseline, which is "no/little importance".

We can interpret these coefficients in terms of odds for each logit equation. For example:

```
exp(1.587703)
```

```
[1] 4.892498
```

is the odds multiplier when comparing "important" versus "no/little importance" for age group >40 compared to age group 18-23. The positive coefficient (or greater than 1 odds multiplier) tells us that older people are more likely to consider the features important than young people, which is consistent with what we observed in the exploratory plots. The precise interpretation of the odds ratio is as follows: The odds of considering the features important (versus "no/little importance") for over 40 year-olds are 4.89 times the odds for 18-23 year olds.

Similarly, we can calculate the odds multiplier for comparing "important" versus "no/little importance" for age group 24-40 compared to age group 18-23:

```
exp(1.128266)
```

```
[1] 3.090293
```

and the positive coefficient also indicates these odds are larger than the baseline category.

The positive coefficient for women (corresponding to an odds multiplier greater than 1):

```
exp(0.3881269)
```

```
[1] 1.474217
```

indicate that women are more likely than men to consider the features important.

We can calculate approximate 95% confidence intervals for these odds multipliers to see which of these differences are actually significant.

```
exp(c(0.3881269-1.96*0.3005110,  0.3881269+1.96*0.3005110))
```

```
[1] 0.818015 2.656816
```

In the table below we have a summary of the coefficients, odds ratios and confidence intervals for the logit equation corresponding to "important" versus "no/little importance". The odds ratio comparing women to men is not significant, but all the odds ratios comparing age groups are.

Parameter, β	Estimate, $\hat{\beta}$ (std. error)	$OR = e^{\hat{\beta}}$ (95% confidence interval)	
$\log(p_2/p_1)$: important vs. no/little importance			
β_{02}: constant	-0.979 (0.256)		
β_{12}: women	0.388 (0.301)	1.47	(0.82, 2.66)
β_{22}: 24-40	1.128 (0.342)	3.09	(1.58, 6.04)
β_{32}: >40	1.590 (0.403)	4.90	(2.22, 10.78)

[a] Source: McFadden, M. J. Powers, W. Brown, and M. Walker (2000). Vehicle and driver attributes affecting distance from the steering wheel in motor vehicles, *Human Factors 42*, 676-682.

Task 1.

Fill in the table below with the relevant odds multipliers and 95% confidence intervals for the "very important" versus "no/little importance" logit equation.

Parameter, β	Estimate, $\hat{\beta}$ (std. error)	$OR = e^{\hat{\beta}}$ (95% confidence interval)	
$\log(p_3/p_1)$: very important vs. no/little importance			
β_{03}: constant	-1.852 (0.331)		
β_{13}: women	0.813 (0.321)	?	(?, ?)
β_{23}: 24-40	1.478 (0.401)	?	(?, ?)
β_{33}: >40	2.917 (0.423)	?	(?, ?)

Model comparisons for the car preference data

We can compare the nominal logistic regression model with additive terms for age and sex with the null model by taking the difference in deviances (likelihood ratio test).

The null model can be fit as follows:

```
nullm <- multinom(response ~ 1, data=dcars, weights=frequency)
```

```
# weights:  6 (2 variable)
initial  value 329.583687
final  value 329.272024
converged
```

```
summary(nullm)
```

```
Call:
multinom(formula = response ~ 1, data = dcars, weights = frequency)

Coefficients:
               (Intercept)
important      -0.11066559
very important -0.03883986

Std. Errors:
               (Intercept)
important        0.1419933
very important   0.1393729

Residual Deviance: 658.544
AIC: 662.544
```

The difference in deviance is $658.54 - 580.70 = 77.84$ which is significant when compared with a $\chi^2(8-2)$:

```
qchisq(df=6, p=0.95)
```

```
[1] 12.59159
```

Overall, the explanatory variables are descriptive of car preferences.

We can also compare this model with the saturated (full) model, which includes an interaction between age and sex:

```
m2 <- multinom(response ~ age * sex, weight = frequency, data = dcars)
```

```
# weights:  21 (12 variable)
initial  value 329.583687
iter  10 value 288.541004
final  value 288.381742
converged
```

```
summary(m2)
```

```
Call:
multinom(formula = response ~ age * sex, data = dcars, weights = frequency)

Coefficients:
               (Intercept)  age24-40  age> 40   sexwomen age24-40:sexwomen age> 40:sexwomen
important        -0.855691 0.7305883 1.484325 0.08253857         0.8898474        0.3183712
very important   -1.609489 1.2612152 2.420439 0.29736599         0.5616963        0.9957248
```

```
Std. Errors:
               (Intercept)  age24-40   age> 40  sexwomen age24-40:sexwomen age> 40:sexwomen
important        0.2895236 0.4575083 0.5248736 0.4534513         0.6998827        0.8177397
very important   0.3873041 0.5405208 0.5749454 0.5756020         0.8070024        0.8580499

Residual Deviance: 576.7635
AIC: 600.7635
```

The difference in deviance between the additive and the saturated model is $580.70 - 576.76 = 3.94$. This is not significant when compared with a $\chi^2(12-8)$, so the additive model appears to fit the data well.

```
qchisq(df=4, p=0.95)
```

```
[1] 9.487729
```

The same conclusion is supported when comparing the AIC for these models: the additive model has a smaller AIC of 596.7 compared to the interaction model which has AIC of 600.7.

Task 2.

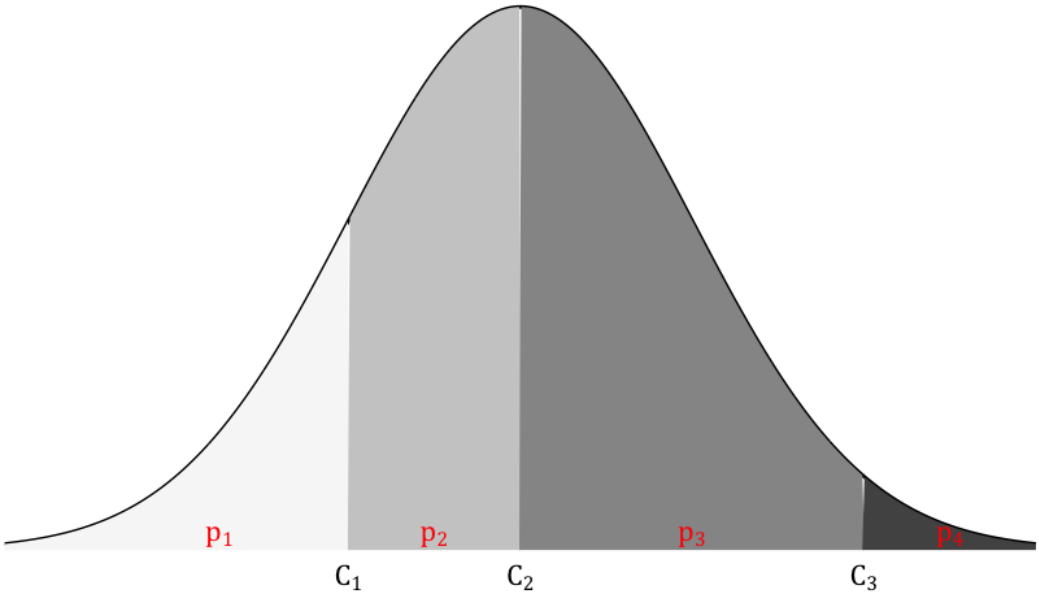
Fit a nominal logistic regression model with a linear term for `age` (create a new `agelin` variable taking values 0, 1 and 2 corresponding to 18-23, 24-40 and >40). Compare this nominal logistic regression model to the model with `age` as a categorical predictor. Which model would you choose and why?

Ordinal logistic regression

In the car preferences example there was a natural ordering among the response categories for the importance of power steering and air conditioning when buying a car: "no/little importance", "important", "very important". This ordering can be taken into account in the model specification. Such ordering often arises in market research, opinion polls and questionnaires (*e.g.* student feedback at the University of Glasgow).

Latent variable view of ordered responses

Sometimes an ordinal response could arise if there is a continuous variable Z, such as severity of disease, which is hard to measure. Z is a **latent variable**, because it cannot be observed directly. Instead, cutpoints C_j are identified so that, for instance, patients have "no disease", "mild disease", "moderate disease" or "severe disease" corresponding to values of Z from low to high. $C_1, \ldots, C_{J-1}$ identify J ordered categories with associated probabilities $p_1, p_2, \ldots, p_J$. An example of the continuous distribution of Z with cutpoints for four categories is shown below. Here, four discrete responses can occur depending on the position of Z relative to the cutpoints C_j.

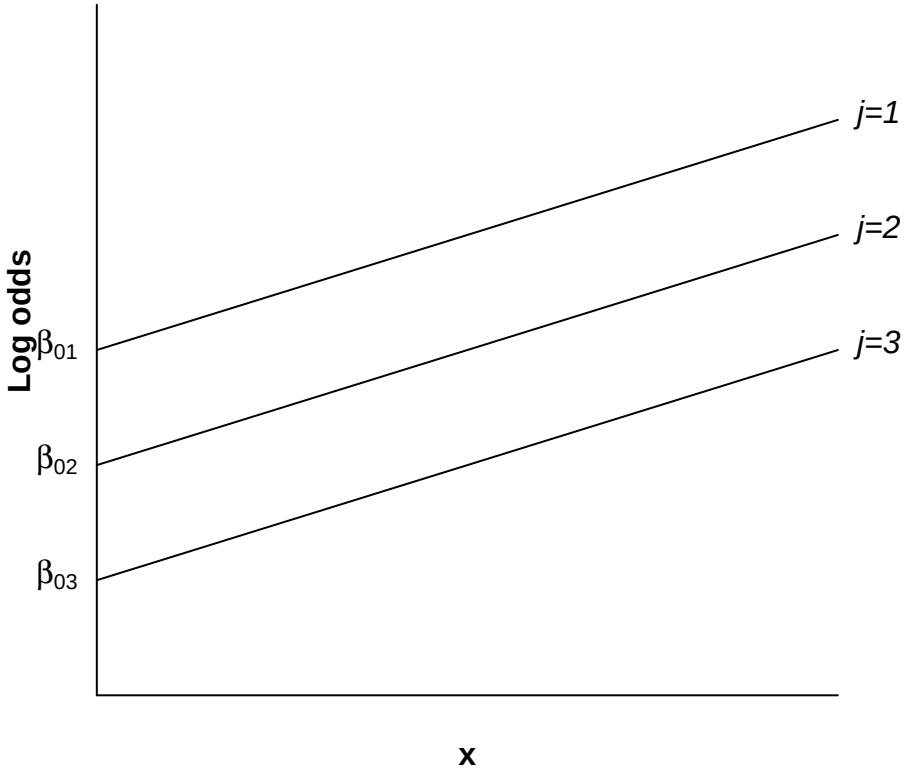


Proportional odds logistic regression model

There are several ways in which to model logits involving the probabilities p_j. The most commonly used model is the **proportional odds logistic regression model**. If the linear predictor $\boldsymbol{x}^\top \boldsymbol{\beta}_j$ has an intercept term β_{0j} which depends on category j, but the other explanatory variables do not depend on j, then the model is

$$\log\left(\frac{p_1 + p_2 + \cdots + p_j}{p_{j+1} + \cdots + p_J}\right) = \beta_{0j} + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}.$$

This is called the **proportional odds model** and is based on the assumption that the effects of the covariates $x_1, \ldots, x_{p-1}$ are the same for all categories on the logarithmic scale, as illustrated in the figure below.



Supplementary material:

Some alternatives to the proportional odds model for ordinal responses are given below.

- Cumulative logit model

The cumulative odds for the jth category are

$$\frac{\Pr(Z \leq C_j)}{\Pr(Z > C_j)} = \frac{p_1 + p_2 + \cdots + p_j}{p_{j+1} + \cdots + p_J}$$

The cumulative logit model is

$$\log\left(\frac{p_1 + p_2 + \cdots + p_j}{p_{j+1} + \cdots + p_J}\right) = \boldsymbol{x}^\top \boldsymbol{\beta}_j.$$

- Adjacent categories logit model

If we consider ratios of probabilities, *e.g.* $\frac{p_1}{p_2}, \frac{p_2}{p_3}, \ldots, \frac{p_{J-1}}{p_J}$ we can define the adjacent category logit model as

$$\log\left(\frac{p_j}{p_{j+1}}\right) = \boldsymbol{x}^\top \boldsymbol{\beta}_j, \qquad \text{for } j = 1, \ldots, J-1.$$

If this is simplified to

$$\log\left(\frac{p_j}{p_{j+1}}\right) = \beta_{0j} + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1},$$

the effect of each explanatory variable is assumed to be the same for all adjacent pairs of categories.

- Continuation ratio logit model

Another alternative is to consider the ratios of probabilities $\frac{p_1}{p_2}, \frac{p_1+p_2}{p_3}, \ldots, \frac{p_1+\cdots+p_{J-1}}{p_J}$ or $\frac{p_1}{p_2+\cdots+p_J}, \frac{p_2}{p_3+\cdots+p_J}, \ldots, \frac{p_{J-1}}{p_J}$.

The equation

$$\log\left(\frac{p_j}{p_{j+1}+\cdots+p_J}\right) = \boldsymbol{x}^\top \boldsymbol{\beta}_j$$

models the odds of the response being in category j, i.e. $C_{j-1} < Z \leq C_j$ conditional upon $Z > C_{j-1}$.

For instance, in the car preferences data example we could estimate the odds of respondents regarding air conditioning and power steering as "unimportant" vs. "important" or "very important" using

$$\log\left(\frac{p_1}{p_2+p_3}\right).$$

Similarly, the odds of these features being "very important" given that they are "important" or "very important" can be estimated by

$$\log\left(\frac{p_2}{p_3}\right).$$

Ordinal logistic regression

https://youtu.be/d9DSd33sUhw

Duration: 5m35s

Example 2 (Proportional odds logistic regression model for the car preference data).

Looking at the car preference example again, we can fit the response as an ordinal variable using a proportional odds model of the form:

$$\log\left(\frac{p_1}{p_2+p_3}\right) = \beta_{01} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{5}$$

$$\log\left(\frac{p_1+p_2}{p_3}\right) = \beta_{02} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{6}$$

where $j = 1$ for "no/little importance" (also referred to as "unimportant"), $j = 2$ for "important" and $j = 3$ for "very important", $x_1 = 1$ for women and 0 for men, $x_2 = 1$ for age 24-40 years and 0 otherwise and $x_3 = 1$ for age > 40 and 0 otherwise.

We fit this model using the `polr()` function in `library(MASS)`, which, incidentally, uses the parameterisation

$$\log\left(\frac{p_1}{p_2+p_3}\right) = \beta_{01} - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 \tag{7}$$

$$\log\left(\frac{p_1+p_2}{p_3}\right) = \beta_{02} - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 \tag{8}$$

instead of (5) and (6).

```
library(MASS)
m4 <- polr(response ~ sex + age, data= dcars, weight = frequency, Hess=TRUE)
summary(m4)
```

```
Call:
polr(formula = response ~ sex + age, data = dcars, weights = frequency,
    Hess = TRUE)

Coefficients:
          Value Std. Error t value
sexwomen 0.5762     0.2262   2.548
age24-40 1.1471     0.2776   4.132
age> 40  2.2325     0.2915   7.659

Intercepts:
                           Value  Std. Error t value
no/little|important        0.6198 0.2168     2.8588
important|very important 2.2312 0.2546     8.7625

Residual Deviance: 581.2956
AIC: 591.2956
```

The intercepts correspond to the jth category, so the log odds of considering the features "unimportant" is 0.620 corresponding to a probability of 0.65 for men age 18-23. The log odds of considering the features "unimportant" or "important" is 2.231 corresponding to a probability of 0.903, giving a probability of 0.253 of considering the features "important". This leaves a probability of 0.097 of considering the features "very

important" for men age 18-23. These probabilities are calculated using equations (7) and (8) together with $p_1 + p_2 + p_3 = 1$. For instance for men age 18-23 (baseline) we get

$$\hat{p}_1 = \frac{\exp(\hat{\beta}_{01})}{1 + \exp(\hat{\beta}_{01})} = \frac{\exp(0.6198)}{1 + \exp(0.6198)} = 0.650$$

and

$$\hat{p}_3 = \frac{1}{1 + \exp(\hat{\beta}_{02})} = \frac{1}{1 + \exp(2.2312)} = 0.097.$$

The likelihood ratio chi-squared statistic for the proportional odds model is 77.25, and the AIC is 591.3, both very similar to those obtained from the corresponding nominal logistic regression model (77.84 and 596.70 respectively). There is little difference in how well the proportional odds and nominal logistic regression models describe the data.

Task 3.

Fit a proportional odds logistic regression model with age as an ordered variable with values 0, 1 and 2 corresponding to 18-23, 24-40 and >40. Is this model preferable to the model with age as a factor?

Task 4.

The `housing` dataset from `library(MASS)` shows a four-way classification of 1681 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. The response, `Sat`, gives the satisfaction of householders with their present housing circumstances, (High, Medium or Low). The explanatory variables are

- `Infl`, the perceived degree of influence householders have on the management of the property (High, Medium, Low),
- `Type`, the type of rental accommodation (Tower, Atrium, Apartment, Terrace), and
- `Cont`, the contact residents are afforded with other residents (Low, High).

Fit nominal and ordinal logistic regression models to these data and interpret the results.

Finally let us look at an application of nominal logistic regression to a classification problem.

Nominal logistic regression as a classification tool

We can use the predicted probabilities from a nominal logistic regression model to classify an observation to the category with the highest predicted probability.

Example 3 (Iris data).

This is a very famous example used for illustrating various classification methods. The data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of three species of iris: setosa, versicolor, and virginica. Suppose that we have these measurements from an iris and we wish to classify it into one of the three species. We can fit a nominal logistic regression model to the data and predict the probability of each species from the fitted model. Let us start with a model with only sepal length as the predictor.

```
library(nnet)
m.iris <- multinom(Species ~ Sepal.Length, data=iris)

# weights:  9 (4 variable)
initial  value 164.791843
```

```
iter  10 value 91.337114
iter  20 value 91.035008
final  value 91.033971
converged
```

```
summary(m.iris)
```

```
Call:
multinom(formula = Species ~ Sepal.Length, data = iris)

Coefficients:
           (Intercept) Sepal.Length
versicolor   -26.08339     4.816072
virginica    -38.76786     6.847957

Std. Errors:
           (Intercept) Sepal.Length
versicolor    4.889635    0.9069211
virginica     5.691596    1.0223867

Residual Deviance: 182.0679
AIC: 190.0679
```

We can get predicted probabilities from this model by using `fitted(m.iris)`. Here are the first few values:

```
head(round(fitted(m.iris),3))
```

```
  setosa versicolor virginica
1  0.807      0.176     0.017
2  0.918      0.077     0.005
3  0.968      0.031     0.001
4  0.980      0.019     0.001
5  0.873      0.118     0.009
6  0.478      0.442     0.080
```

And if we use `predict(m.iris)`, we get a list which assigns each observation to the category with the highest predicted probability, for instance the first observation is assigned to `setosa` with predicted probability = 0.807. Here are the first few values:

```
head(predict(m.iris))
```

```
[1] setosa setosa setosa setosa setosa setosa
Levels: setosa versicolor virginica
```

We can see how many predictions the model got wrong:

```
sum(iris$Species!=predict(m.iris))
```

```
[1] 38
```

Of course, if we were to properly assess the classification performance of this model, we should look at out-of-sample prediction by first splitting the data into a training and a test set, then fitting the model to the training data and finally predicting the class of the each observation in the test data. Otherwise we run the risk of overstating the classification accuracy of the model.

Task 5.

Fit a nominal logistic regression model to the iris data using all four predictors. Do you get better classification performance?

More examples and details on GLMs for nominal and ordinal data can be found in

- Chapter 5 from **Extending linear models with R: generalized linear, mixed effects and nonparametric regression models by Julian Faraway** and in
- Chapter 6 of **Regression: models, methods and applications by Fahrmeir et al.**

R examples are also available from UCLA's Institute for Digital Research and Education:

- Nominal logistic regression example
- Ordinal logistic regression example

Week 4 learning outcomes

- Identify categorical responses as nominal or ordinal
- Fit nominal logistic regression models to both nominal and ordinal data using the `multinom()` function in `library(nnet)`
- Fit ordinal logistic regression models for ordered responses using the `polr()` function in `library(MASS)`
- Choose a model by comparing deviances and/or AIC
- Interpret model coefficients in terms of odds ratios
- Obtain predicted probabilities and fitted values from a nominal logistic regression model or a proportional odds model
- Use nominal logistic regression as a classification tool

Answers to tasks

Answer to Task 1. We calculate the odds ratios and corressponding 95% CI as follows:

```
# women Odds Ratio
exp(0.813)
```

```
[1] 2.254662
```

```
# women 95% CI Odds Ratio
exp(c(0.813-1.96*0.321, 0.813+1.96*0.321))
```

```
[1] 1.201824 4.229822
```

```
# 24-40 Odds Ratio
exp(1.478)
```

```
[1] 4.384169
```

```
# 24-40 95% CI Odds Ratio
exp(c(1.478-1.96*0.401, 1.478+1.96*0.401))
```

```
[1] 1.997787 9.621113
```

```
# >40 Odds Ratio
exp(2.917)
```

```
[1] 18.48575
```

```
# >40 95% CI Odds Ratio
exp(c(2.917-1.96*0.423, 2.917+1.96*0.423))
```

```
[1]  8.068116 42.354726
```

Parameter, β	Estimate, $\hat{\beta}$ (std. error)	$OR = e^{\hat{\beta}}$ (95% confidence interval)	
$\log(p_3/p_1)$: very important vs. no/little importance			
β_{03}: constant	-1.852 (0.331)		
β_{13}: women	0.813 (0.321)	2.25	(1.20, 4.23)
β_{23}: 24-40	1.439 (0.401)	4.38	(2.00, 9.62)
β_{33}: >40	2.917 (0.423)	18.48	(8.07, 42.34)

Answer to Task 2. First, construct a linear term for age:

```
dcars$agelin <- 0
dcars$agelin[dcars$age=="24-40"] <- 1
dcars$agelin[dcars$age=="> 40"] <- 2
```

Then fit the model with this new term:

```
m3 <- multinom(response ~ agelin + sex, weight = frequency, data = dcars)
```

```
# weights:  12 (6 variable)
initial  value 329.583687
iter  10 value 291.050616
final  value 291.050160
converged
```

```
summary(m3)
```

```
Call:
multinom(formula = response ~ agelin + sex, data = dcars, weights = frequency)

Coefficients:
               (Intercept)    agelin  sexwomen
important       -0.8908983 0.8303799 0.3889732
very important  -1.9053456 1.5214463 0.8130386

Std. Errors:
               (Intercept)    agelin  sexwomen
```

```
important        0.2402679 0.1946354 0.2991328
very important   0.3089657 0.2114578 0.3211055

Residual Deviance: 582.1003
AIC: 594.1003
```

The model with a linear term for age fits the data almost as well as that with age as a factor, and has two fewer parameters. As the models are nested, we can use the difference deviance between them to choose a model:

```
m3$deviance-m1$deviance
```

```
[1] 1.398123
```

```
qchisq(df=2, p=0.95)
```

```
[1] 5.991465
```

As $1.40 < 5.99$, we can go with the simpler model with the linear term in age. This assumes that the odds multiplier is the same when comparing 24-40 year olds to 18-23 year olds and over 40s to 24-40 year olds.

Answer to Task 3. Model with a linear term for age:

```
m5 <- polr(response ~ sex + agelin, data= dcars, weight = frequency, Hess=TRUE)
summary(m5)
```

```
Call:
polr(formula = response ~ sex + agelin, data = dcars, weights = frequency,
    Hess = TRUE)

Coefficients:
         Value Std. Error t value
sexwomen 0.577     0.2261   2.552
agelin   1.116     0.1457   7.660

Intercepts:
                          Value  Std. Error t value
no/little|important       0.6101 0.2034     2.9999
important|very important 2.2214 0.2430     9.1426

Residual Deviance: 581.3124
AIC: 589.3124
```

The parameter estimates from the proportional odds model are very similar to those of the nominal regression model, whether age is included as a factor or as an ordered variable. Fitted probabilities for each covariate pattern are also very similar. The proportional odds model would be preferred because it fits the data as well as the nominal regression model but uses fewer parameters, and because it takes into account the ordinal nature of the response.

Answer to Task 4. First fit a nominal logistic regression model:

```
library(MASS)
library(nnet)
house.nom <- multinom(Sat ~ Infl + Type + Cont, weights = Freq, data = housing)
```

```
# weights:  24 (14 variable)
initial  value 1846.767257
iter  10 value 1747.045232
final  value 1735.041933
converged
```

```
summary(house.nom, digits=3)
```

```
Call:
multinom(formula = Sat ~ Infl + Type + Cont, data = housing,
    weights = Freq)

Coefficients:
```

```
       (Intercept) InflMedium InflHigh TypeApartment TypeAtrium TypeTerrace ContHigh
Medium      -0.419      0.446    0.665        -0.436      0.131      -0.667    0.361
High        -0.139      0.735    1.613        -0.736     -0.408      -1.412    0.482

Std. Errors:
       (Intercept) InflMedium InflHigh TypeApartment TypeAtrium TypeTerrace ContHigh
Medium       0.173      0.142    0.186         0.173      0.223       0.206    0.132
High         0.159      0.137    0.167         0.155      0.211       0.200    0.124

Residual Deviance: 3470.084
AIC: 3498.084
```

The baseline type of housing used for comparisons is tower block, with low influence on apartment management and low contact with other residents. The positive coefficients of `Infl` (medium and high) and `Cont` indicate that satisfaction increases with the feeling of influence and with more contact with other residents. Other types of housing are associated with lower satisfaction ratings than tower block (with the exception of atrium which has a non-significant coefficient).

Given the ordinal nature of the response, we can fit a proportional odds model to see if it fits the data just as well as the nominal logistic regression.

```
library(MASS)
house.plr <- polr(Sat ~ Infl+Type+Cont, weights=Freq, data=housing, Hess=TRUE)
summary(house.plr, digits = 3)
```

```
Call:
polr(formula = Sat ~ Infl + Type + Cont, data = housing, weights = Freq,
    Hess = TRUE)

Coefficients:
               Value Std. Error t value
InflMedium     0.566     0.1047    5.41
InflHigh       1.289     0.1272   10.14
TypeApartment -0.572     0.1192   -4.80
TypeAtrium    -0.366     0.1552   -2.36
TypeTerrace   -1.091     0.1515   -7.20
ContHigh       0.360     0.0955    3.77

Intercepts:
            Value   Std. Error t value
Low|Medium  -0.496  0.125     -3.974
Medium|High  0.691  0.125      5.505

Residual Deviance: 3479.149
AIC: 3495.149
```

The results obtained from a proportional odds regression are qualitatively similar: in general satisfaction increases with influence and contact. Tower block seems to be the type of housing with the highest satisfaction ratings, followed by atrium, apartment and terrace. The proportional odds model is simpler and takes into account the ordinal nature of the response variable.

The following code produces approximate confidence intervals for the coefficients of the proportional odds model:

```
confint(house.plr)
```

```
                    2.5 %      97.5 %
InflMedium     0.3616417  0.77195442
InflHigh       1.0409711  1.53958312
TypeApartment -0.8069602 -0.33940475
TypeAtrium    -0.6705869 -0.06204459
TypeTerrace   -1.3893882 -0.79534035
ContHigh       0.1733591  0.54792915
```

We can convert to odds ratios by exponentiating:

```
round(exp(confint(house.plr)),2)
```

```
                2.5 % 97.5 %
InflMedium      1.44   2.16
InflHigh        2.83   4.66
TypeApartment   0.45   0.71
TypeAtrium      0.51   0.94
TypeTerrace     0.25   0.45
ContHigh        1.19   1.73
```

Interpretation for the Influence coefficients: The odds of higher satisfaction (low to medium/medium to high) are between 1.44 and 2.16 times higher for medium influence compared to low influence, and between 2.83 and 4.66 times higher for high influence compared to low influence. The other coefficients are interpreted similarly.

Answer to Task 5. Without increasing the maximum number of iterations you would get a warning about lack of convergence.

```
m.iris.all <- multinom(Species ~ . , data=iris)
```

```
# weights:  18 (10 variable)
initial  value 164.791843
iter  10 value 16.177348
iter  20 value 7.111438
iter  30 value 6.182999
iter  40 value 5.984028
iter  50 value 5.961278
iter  60 value 5.954900
iter  70 value 5.951851
iter  80 value 5.950343
iter  90 value 5.949904
iter 100 value 5.949867
final  value 5.949867
stopped after 100 iterations
```

The maximum number of iterations can be increased using `maxit`:

```
m.iris.all <- multinom(Species ~ . , data=iris, maxit=1000)
```

```
# weights:  18 (10 variable)
initial  value 164.791843
iter  10 value 16.177348
iter  20 value 7.111438
iter  30 value 6.182999
iter  40 value 5.984028
iter  50 value 5.961278
iter  60 value 5.954900
iter  70 value 5.951851
iter  80 value 5.950343
iter  90 value 5.949904
iter 100 value 5.949867
iter 110 value 5.949850
iter 120 value 5.949821
iter 130 value 5.949767
iter 140 value 5.949743
iter 150 value 5.949722
iter 160 value 5.949686
iter 170 value 5.949424
iter 180 value 5.949393
final  value 5.949363
converged
```

```
summary(m.iris.all)
```

```
Call:
multinom(formula = Species ~ ., data = iris, maxit = 1000)
```

```
Coefficients:
           (Intercept) Sepal.Length Sepal.Width Petal.Length Petal.Width
versicolor    18.40821    -6.082250   -9.396625     16.17037   -2.058115
virginica    -24.23006    -8.547304  -16.077164     25.59963   16.227474

Std. Errors:
           (Intercept) Sepal.Length Sepal.Width Petal.Length Petal.Width
versicolor    22.60419     38.59747    40.37282     109.0391    60.45113
virginica     23.61506     38.61518    40.53550     109.1776    60.76801

Residual Deviance: 11.89873
AIC: 31.89873
```

Notice the large standard errors in the model output. These point to perfect prediction/separation. Check:

```
sum(iris$Species!=predict(m.iris.all))
```

```
[1] 2
```

All but two observations appear to have been predicted correctly.

The function `glmnet()` might be helpful for dealing with separation.