STAT3015/7030: Generalised Linear Modelling
Multinomial Models

Bronwyn Loong

Semester 2 2014

References

Ch 6.5 - Gelman and Hill

Ch 5 - Faraway

Multinomial distribution

- Consider data that has J categories: $\{1, \ldots, J\}$.
- These categories could be:
  - unordered: e.g. eye color (brown, green, blue, gray);
  - ordered: e.g. company rankings (analyst, associate, director, partner).
- Let $Y_{i,j}$ be the number of observations falling into category j for group i.

$$Pr(Y_{i,1} = y_{i,1}, \ldots, Y_{i,J} = y_{i,J}) = \frac{n_i!}{y_{i,1}! \cdots y_{i,J}!} \pi_{i,1}^{y_{i,1}} \cdots \pi_{i,J}^{y_{i,J}}$$

  where $n_i = \sum_j y_{ij}$ and $\sum_j \pi_{i,j} = 1$. So we have a multinomial distribution!
- Note that if we have ungrouped data where $n_i = 1$ then:

$$p(Y_{i,1} = y_{i,1}, \ldots, Y_{i,J} = y_{i,J}) = \pi_{i,1}^{y_{i,1}} \cdots \pi_{i,J}^{y_{i,J}}$$

Unordered Categories

As usual, we need to find a way to link our covariates with the mean (or probability) for the one of the j categories. A possibility is:

$$log\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = \eta_{ij} = x_i^T\beta_j \quad \text{for } j = 2, \ldots, J$$

So we have:

$$\pi_{i,j} = exp(\eta_{i,j})\pi_{i,1}$$

Notice that we compare the $2, \ldots, J$ categories to the first category! Similar to the notion for factors under treatment coding! Additionally, we must remember $\sum_j \pi_{i,j} = 1$. So we set:

$$\pi_{i,1} = 1 - \sum_{j=2}^{J} \pi_{i,j}$$

Unordered Categories

So then:

$$\begin{aligned}\pi_{i,1} &= 1 - \sum_{j=2}^{J} exp(\eta_{i,j})\pi_{i,1} \\ &= \frac{1}{(1 + \sum_{j=2}^{J} exp(\eta_{i,j}))}\end{aligned}$$

Which leads to:

$$\pi_{i,j} = \frac{exp(\eta_{i,j})}{1 + \sum_{j=2}^{J} exp(\eta_{i,j})}$$

Based on this formulation, a likelihood can be formed and parameter estimation can be conducted via maximum likelihood estimation and then we can use our standard methods for inference.

Example - Unordered Categories

Consider the 1996 American Election Study data. We will only consider a few covariates and collapse the categories into a three:

- Categories = {Democrat, Independent, Republican};
- age: Respondent's age in years
- educ: Respondent's education: an ordered factor with levels 8 years or less
- income: Respondent's family income: an ordered factor with levels $3Kminus ; $3K-$5K ; $5K-$7K ; $7K-$9K ; $9K-$10K ; $10K-$11K ; $11K-$12K ; $12K-$13K ; $13K-$14K ;$14K-$15K ; $15K-$17K ; $17K-$20K ; $20K-$22K ; $22K-$25K ; $25K-$30K ; $30K-$35K ; $35K-$40K ; $40K-$45K ; $45K-$50K ; $50K-$60K ; $60K-$75K ; $75K-$90K ; $90K-$105K ; $105Kplus.

Notice that the data are not grouped. We have information for each individual. What does this mean for using the deviance as a goodness-of-fit statistic?

Example - Unordered Categories

```
> library(faraway)
> ## data
> data(nes96)
> sPID <-  nes96$PID
> ## collapse categories
> levels(sPID) <- c("D","D", "I", "I", "I", "R", "R")
> ## treat income as continuous
> inca <- c(1.5, 4, 6, 8, 9.5, 10.5, 11.5, 12.5, 13.5, 14.5
+   16, 18.5, 21, 23.5, 27.5, 32.5,
+   37.5, 42.5, 47.5, 55, 67.5, 82.5,
+   97.5, 115)
#unclass - convert factor to integer codes
> income <- inca[unclass(nes96$income)]
> educ <- as.factor(unclass(nes96$educ))
> age <- nes96$age
```

Example - Unordered Categories

```
> head(data.frame(sPID, income, educ, age))
  sPID income educ age
1    R    1.5    3  36
2    D    1.5    4  20
3    D    1.5    6  24
4    D    1.5    6  28
5    D    1.5    6  68
6    D    1.5    4  21
> summary(income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.50   23.50   37.50   46.58   67.50  115.00
> table(educ)
educ
  1   2   3   4   5   6   7
 13  52 248 187  90 227 127
```

Example - Unordered Categories

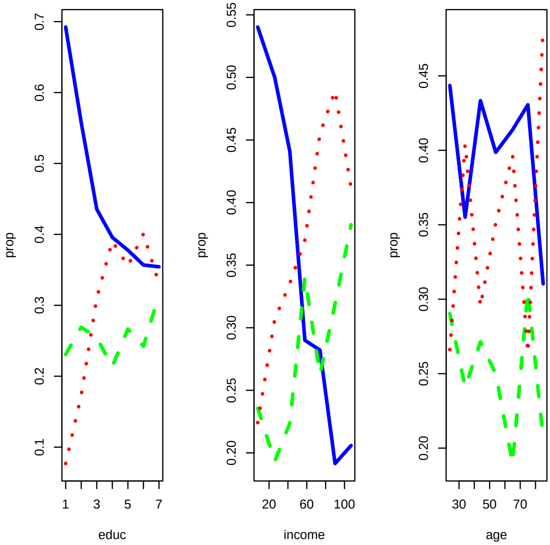


Figure : Empirical probabilities vs covariates (broken into categories) - D (blue), I (green), R (red)

Example - Unordered Categories

```
#let's fit a model
> library(nnet)
> mmod <- multinom(sPID ~ age + educ + income)
# weights:  30 (18 variable)
initial  value 1037.090001
iter  10 value 990.364722
iter  20 value 984.508641
final  value 984.166272
converged
> summary(mmod)
Call:
multinom(formula = sPID ~ age + educ + income)

Coefficients:
  (Intercept)          age     educ2     educ3      educ4
I   -1.373895 0.0001539014 0.2704482 0.2458744 0.09119446
R   -3.048576 0.0081945031 0.9876547 1.6915600 1.95336096
      educ5     educ6     educ7     income
I 0.3269554 0.1082654 0.1933497 0.01623914
R 1.8835335 1.8708213 1.4539589 0.01724696
```

Example - Unordered Categories

```
Std. Errors:
  (Intercept)         age     educ2     educ3    educ4     educ5
I    0.766464 0.005374592 0.7460643 0.6992364 0.713322 0.7382877
R    1.111205 0.004902674 1.1237993 1.0721857 1.076931 1.0940829
      educ6     educ7      income
I 0.7151263 0.7287344 0.003108590
R 1.0792352 1.0914118 0.002881749

Residual Deviance: 1968.333
AIC: 2004.333
```

Example - Unordered Categories

How do we interpret the βs? Remember we are comparing Independent to Democrat & Republican to Democrat. This is just the same as the logistic regression interpretation! Let's consider age.
Consider the following:

$$log\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = log \text{ odds } = \beta_j' x_i = \beta_{0,j} + \beta_{1,j} x_{age,i} + \cdots$$

Now let's increase x_{age} by one unit (dropping the i for convenience):

$$\begin{aligned} log \text{ odds } (x_{age}\uparrow) - log \text{ odds } (x_{age}) &= \beta_{0,j} + \beta_{1,j} x_{age}(\uparrow) + \cdots - \beta_{0,j} - \beta_{1,j} x_{age} - \cdots \\ &= \beta_{1,j} x_{age}(\uparrow) - \beta_{1,j} x_{age} \\ &= \beta_{1,j}(x_{age}(\uparrow) - x_{age}) \\ &= \beta_{1,j} \times 1 \end{aligned}$$

Example - Unordered Categories

What about interpretation in relation to the odds?

$$\begin{aligned} log \text{ odds } (x_{age} \uparrow) - log \text{ odds } (x_{age}) &= log\left(\frac{\text{odds } (x_{age} \uparrow)}{\text{odds } (x_{age})}\right) \\ log\left(\frac{\text{odds } (x_{age} \uparrow)}{\text{odds } (x_{age})}\right) &= \beta_{1,j} \\ \left(\frac{\text{odds } (x_{age} \uparrow)}{\text{odds } (x_{age})}\right) &= exp(\beta_{1,j}) \end{aligned}$$

or

$$\text{odds } (x_{age} \uparrow) = \text{ odds } (x_{age}) exp(\beta_{1,j})$$

So if we increase age by one year:

- the odds of being an independent compared to a democrat changes by a factor of:

```
> exp(0.000154)
[1] 1.00015
```

- the odds of being a republican compared to a democrat changes by a factor of:

```
> exp(0.0081945)
[1] 1.00823
```

So not much change! This agrees with our plot!

Example - Unordered Categories

```
##can we reduce the model?

> mmod.aic <- step(mmod, trace=0)

trying - age
trying - educ
trying - income
# weights:  12 (6 variable)
initial  value 1037.090001
iter  10 value 992.269502
final  value 992.269484
converged
trying - age
trying - income
# weights:  9 (4 variable)
initial  value 1037.090001
final  value 992.712152
converged
trying - income
```

Example - Unordered Categories

```
> ## deviance test for dropping educ & age
> diff.dev <- deviance(mmod.aic) - deviance(mmod)
> 1 - pchisq(diff.dev, mmod$edf-mmod.aic$edf)
[1] 0.2513210
> ##
> summary(mmod.aic)

Coefficients:
  (Intercept)     income
I   -1.174933 0.01608683
R   -0.950359 0.01766457

Std. Errors:
  (Intercept)      income
I   0.1536103 0.002849738
R   0.1416859 0.002652532

Residual Deviance: 1985.424
AIC: 1993.424
```

Example - Unordered Categories

```
#let's look at a change of $1000 of income
> pp <- predict(mmod.aic, data.frame(income=c(0,1)),
type="probs")
> pp

          D         I         R
1 0.5898168 0.1821588 0.2280244
2 0.5857064 0.1838228 0.2304708

> ## log-odds correspond to slopes
> log( pp[2,2]/pp[2,1]) - log( pp[1,2]/pp[1,1])

[1] 0.01608683

> log( pp[2,3]/pp[2,1]) - log( pp[1,3]/pp[1,1])

[1] 0.01766457
```

You should see that these are the estimated βs!

Example - Unordered Categories

```
> ## Let's predict based on the following incomes in $1,000.
> il <- c(8, 26, 42, 58, 74, 90, 107)
> predict(mmod.aic, data.frame(income=il), type="probs")

          D         I         R
1 0.5566253 0.1955183 0.2478565
2 0.4804946 0.2254595 0.2940459
3 0.4134268 0.2509351 0.3356381
4 0.3493884 0.2743178 0.3762939
5 0.2903271 0.2948600 0.4148129
6 0.2375755 0.3121136 0.4503109
7 0.1891684 0.3266848 0.4841468
```

The probability of being Republican or Independent increases with income.

```
> ## just most probable for each income group
> predict(mmod.aic, data.frame(income=il))
[1] D D D R R R R
Levels: D I R
```

Example - Unordered Categories

```
> ## let''s predict for an income of 0 so just the intercepts!!
> cc <- c(0, -1.174933, -0.950359)
> exp(cc)/(sum(exp(cc)))
[1] 0.5898167 0.1821588 0.2280245

> predict(mmod.aic, data.frame(income=0), type="probs")

        D         I         R
0.5898168 0.1821588 0.2280244
```

Ordered Categories

- It is quite natural, and generally done in the field of Political Science, to consider party affiliations on an ordered scale {D, I, R}.

We are interested in modeling the probabilities:

$$\gamma_{i,j} = p(y_i \leq j) \text{ where } \gamma_{i,J} = 1 \text{ (recall J is the last category)}$$

The cumulative probabilities $\gamma_{i,j}$ are easier to work with. The $\gamma_{i,j}$'s are invariant to combining adjacent categories, Furthermore, we need only model J-1 probabilities.

Ordered Categories

We want to link the γ's to some covariates x.

$$g(\gamma_{i,j}) = \theta_j - x_i'\beta$$

We have explicitly specified the intercepts θ_j so that the vector x_i does not include an intercept.

Also notice that with the ordered case compared to the unordered case, we use less parameters since the βs do not depend on j. That is, we assume the predictors have a uniform effect on the response categories.

Ordered Categories

Latent variable interpretation

It is easier to understand the structure of the multinomial logit model with ordered categories in a latent variable framework

Consider the following model, based on the party identification example:

$$y = \begin{cases} 1 & \text{Democrat;} \\ 2 & \text{Independent;} \\ 3 & \text{Republican.} \end{cases}$$

Now consider z_i to be a continuous latent (unobserved) variable, where y_i is the discretized version of z_i:

$$y_i = \begin{cases} 1 & \text{if} \quad -\infty < z_i < \theta_1; \\ 2 & \text{if} \quad \theta_1 < z_i < \theta_2; \\ 3 & \text{if} \quad \theta_2 < z_i < \infty, \end{cases}$$

Ordered Categories

Now let $z_i - x_i'\beta$ have cumulative distribution function F:

$$\gamma_{i,j} = Pr(y_i \leq j) = Pr(z_i \leq \theta_j) = Pr(z_i - x_i'\beta \leq \theta_j - x_i'\beta) = F(\theta_j - x_i'\beta)$$

- If F follows a logistic distribution:

$$\gamma_{ij} = \frac{exp(\theta_j - x_i'\beta)}{1 + exp(\theta_j - x_i'\beta)}$$

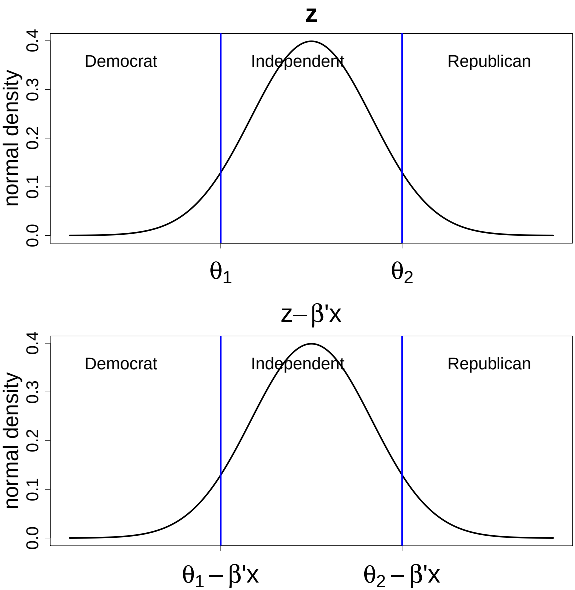
- If F follows a Gaussian distribution (Probit model):

$$\gamma_{ij} = \Phi(\theta_j - x_i'\beta)$$

The figure on the following slide shows the Probit case.

- The first panel shows that the value of z_i and the intercepts θs (cut-points) determine the categories y.
- If we subtract off the mean of z_i, then the covariates become explicit. So if $\beta > 0$ and we increase x we increase the probability of the last category since the blue lines will shift left!

Ordered Categories



Example - Ordered Categories

Let's use the same `nes` data but treat the response categories as ordered. We will use the logit link and fit a proportional odds model

Proportional Odds model

Let $\gamma_{i,j} = P(Y_i \leq j | x_i)$, then the proportional odds model, which uses the logit link, is:

$$\log \frac{\gamma_{i,j}(x_i)}{1 - \gamma_{i,j}(x_i)} = \theta_j - {x_i}^T \beta$$

Here we assume $z_i - x_i' \beta$ follows the density function of a logistic distribution.

It is so called because the relative odds for $y \leq j$ comparing x_1 and x_2 are:

$$\left(\frac{\gamma_{1,j}(x_1)}{1 - \gamma_{1,j}(x_1)} \right) / \left(\frac{\gamma_{2,j}(x_2)}{1 - \gamma_{2,j}(x_2)} \right) = \exp(-(x_1 - x_2)^T \beta)$$

This does not depend on j. We need to check the proportional odds assumption for a given data set.

Example - Ordered Categories

Proportional Odds model

Or note that

$$
\begin{aligned}
\log \frac{\gamma_{i,1}(x_i)}{1-\gamma_{i,1}(x_i)} - \log \frac{\gamma_{i,2}(x_i)}{1-\gamma_{i,2}(x_i)} &= (\theta_1 = x_i^T \beta) - (\theta_2 - x_I^T \beta) \\
&= (\theta_1 - \theta_2) \ \forall_i
\end{aligned}
$$

Example - Ordered Categories

```
> ## Ordered logit
> library(MASS)
> mod <- polr(sPID ~ age + educ + income)
> ## Variable selection via AIC
> mod.aic <- step(mod, trace=0)
> summary(mod.aic)

Coefficients:
         Value Std. Error t value
income 0.01312   0.001971   6.657

Intercepts:
    Value   Std. Error t value
D|I  0.2091  0.1123     1.8627
I|R  1.2916  0.1201    10.7526

Residual Deviance: 1995.363
AIC: 2001.363
```

Example - Ordered Categories

```
> ## Check via difference in deviance
> dev <- deviance(mod.aic)- deviance(mod)
> df <- mod$edf -  mod.aic$edf
> 1- pchisq(dev, df)

[1] 0.1321517
```

So we cannot reject the model chosen via AIC in comparison to the larger model. Let's examine the model:

Example - Ordered Categories

```
> summary(mod.aic)

Call:
polr(formula = sPID ~ income)

Coefficients:
         Value Std. Error t value
income 0.01312   0.001971   6.657

Intercepts:
    Value   Std. Error t value
D|I  0.2091  0.1123     1.8627
I|R  1.2916  0.1201    10.7526

Residual Deviance: 1995.363
AIC: 2001.363
```

Let's check to see whether the proportional odds assumption is violated?

Example - Ordered Categories

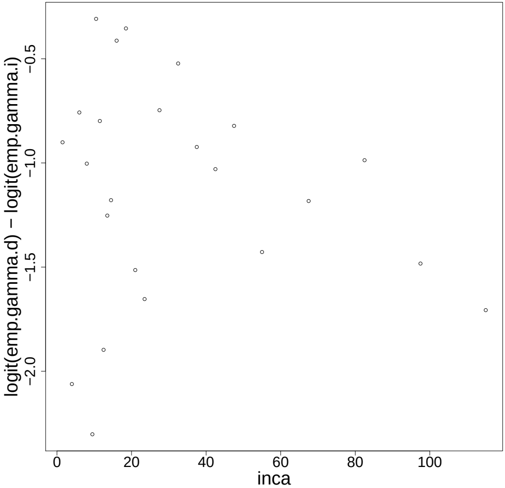
So we examine the empirical differences to see if they are constant!

```
> pim <- prop.table(table(income, sPID),1)
> pim
> emp.gamma.d <- pim[,1]
> emp.gamma.i <- pim[,1]+pim[,2]

> logit(emp.gamma.d) - logit(emp.gamma.i)

      1.5         4         6         8       9.5
-0.9007865 -2.0614230 -0.7576857 -1.0033021 -2.3025851
     10.5      11.5      12.5      13.5      14.5
-0.3083014 -0.7985077 -1.8971200 -1.2527630 -1.1786550
       16      18.5        21      23.5      27.5
-0.4128452 -0.3542428 -1.5141277 -1.6534548 -0.7467847
     32.5      37.5      42.5      47.5        55
-0.5225217 -0.9232594 -1.0296194 -0.8219801 -1.4276009
     67.5      82.5      97.5       115
-1.1826099 -0.9867640 -1.4829212 -1.7066017
```

Example - Ordered Categories



Does not quite look constant, but also doesn't appear to have a trend!

Example - Ordered Categories

Now consider the interpretation of the fitted coefficients:

```
> summary(mod.aic)

Coefficients:
         Value Std. Error t value
income 0.01312   0.001971   6.657

Intercepts:
    Value   Std. Error t value
D|I  0.2091  0.1123     1.8627
I|R  1.2916  0.1201    10.7526
```

The odds of moving from Democrat to Independent/Republican categories (or from Democrat/Independent to Republican) increase by a factor of exp(0.013120)=1.0132 as income increases by one unit ($1000). Notice that the log odds are similar to those obtained in the multinomial logit model.

Example - Ordered Categories

The intercepts correspond to the θ_j, equivalent to the predicted probabilities if the income is 0.

```
> x <- 0
>  gamma.d <- ilogit(0.2091 - mod.aic$coef*x)
>  gamma.i <- ilogit(1.2916  - mod.aic$coef*x)
>  gamma.r <- 1
>  prob.d <- gamma.d
>  prob.i <- gamma.i - gamma.d
>  prob.r <- gamma.r - gamma.i
>  prob.d
   income
0.5520854
> prob.i
   income
0.2323325
> prob.r
   income
0.2155821
```

Example - Ordered Categories

Compute predicted probabilities of Dem/Ind/Rep at different levels of income

```
> predict(mod.aic,data.frame(income=il,row.names=il),
type="probs")
            D         I         R
8   0.5260129 0.2401191 0.2338679
26  0.4670450 0.2541588 0.2787962
42  0.4153410 0.2617693 0.3228897
58  0.3654362 0.2641882 0.3703756
74  0.3182635 0.2612285 0.4205080
90  0.2745456 0.2531189 0.4723355
107 0.2324161 0.2395468 0.5280371
```

Examine the patterns in each party affiliation group across income? How do the patterns differ?

Example - Ordered Categories

Compare cutpoints for incomes of \$0, \$50,000 and \$100,000

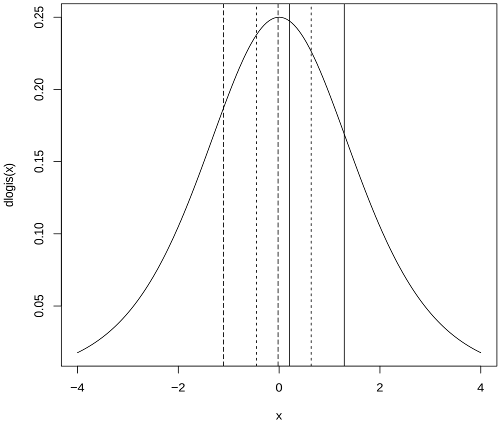


Figure : Solid lines: income=\$0; Dotted lines: income=\$50000; Dashed lines: income=\$100000

Example - Ordered Categories

Ordered Probit Model

Here we assume $z_i - x_i'\beta$ follows the density function of a standard normal distribution.

```
> mod.prob <- polr(formula = sPID ~ income, method="probit")
> summary(mod.prob)

Call:
polr(formula = sPID ~ income, method = "probit")

Coefficients:
          Value Std. Error t value
income 0.008182   0.001208   6.775

Intercepts:
    Value   Std. Error t value
D|I  0.1284  0.0694      1.8510
I|R  0.7976  0.0722     11.0399

Residual Deviance: 1994.892
```

Example - Ordered Categories

Proportional Hazards Model

Concept of a hazard (from insurance)

$$\log(-\log(1-\gamma_j(x_i))) = \theta_j - x_i^T\beta$$

The hazard of category j is the probability of falling in category j given that your category is greater than j (or given that your category doesn't fall in categories 1,...,j-1).

$$\mathrm{Hazard}(j) = Pr(Y_i = j | Y_i \geq j) = \frac{Pr(Y_i = j)}{Pr(Y_i \geq j)} = \frac{\pi_{ij}}{1-\gamma_{i,j-1}} = \frac{\gamma_{i,j}-\gamma_{i,j-1}}{1-\gamma_{i,j-1}}$$

The corresponding latent variable distribution is

$$F(\theta_j - x_i^T\beta) = 1 - \exp(-\exp(\theta_j - x_i^T\beta))$$

Little practical justification to apply it to the `nes96` data

```
mod.cloglog <- polr(formula = sPID ~ income, method="cloglog")
```