IDS 702

Multinomial logistic regression

Model recap

Type of outcome	Model

Multinomial distribution

PMF: $Pr(X_1 = x_1 and \ldots and X_k = x_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}$

Support: $x_i \in \{0,...,n\}, i \in \{1,...,k\}, with \sum_i x_i = n$

Parameters: $n > 0$ number of trials, $k > 0$ number of mutually exclusive events, $p_1, \ldots p_k$ event probabilities ($\sum_i p_i = 1$)

Multinomial distribution practice

- 4 people (Andrea, Sarah, Nick, and Kyle) will play Settlers of Catan this weekend. They will play 5 games. Andrea's probability of winning is 0.65. Sarah's probability of winning is 0.22, Nick's is 0.08, and Kyle's is 0.05. What is the probability that Sarah will win 2 games, Andrea will win 1, Kyle will win 2, and Nick will win 0?
- What if the probabilities of winning are instead 0.26, 0.34, 0.1, and 0.42?

Data example

- A study is conducted to examine factors that may influence a student's choice of career track upon entering high school. Students can choose between 3 tracks: academic, general, vocational
- Potential predictors include: student's socioeconomic status (ses), sex, school type, and various test scores

https://stats.oarc.ucla.edu/r/dae/multinomial-logistic-regression/

https://bookdown.org/chua/ber642_advanced_regression/multinomial-logistic-regression.html

Multinomial distribution

- We can select the multinomial distribution to describe the outcome Y
- $P[y_i = 1 \,|\, x_i] = \pi_{1i}, P[y_i = 2 \,|\, x_i] = \pi_{2i}, \ldots, P[y_i = J \,|\, x_i] = \pi_{Ji}$, where

$$\sum_{j=1}^{J} \pi_{ji} = 1$$

Alternative logistic regression setup

- Recall logistic:
    - $y_i | x_i \sim Bernoulli(\pi_i); log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_i$
    - $\pi_i = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$
- Now consider: $P[y_i = 1 | x_i] = \pi_i = \pi_{1i}$ and $P[y_i = 0 | x_i] = 1 - \pi_i = \pi_{0i}$
- Then we can write the logistic model as $log\left(\frac{\pi_{1i}}{\pi_{0i}}\right) = \beta_0 + \beta_1 x_i$
- And e^{β_1} can be interpreted as the multiplicative change in odds of $y = 1$ over the baseline $y = 0$ when increasing x by one unit

Multinomial logistic regression setup

- We can still use the logit link function and set one level as the baseline, say $Y = 1$ (e.g., academic track)
- Then the multinomial logistic regression is defined as a set of logistic regression models for each probability π_j, compared to the baseline, where $j \geq 2$:

$$log\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = \beta_{0j} + \beta_{1j}x_{i1} + \ldots + \beta_{pj}x_{ip}$$

- So how many separate logistic regression models do we have in terms of J?

Interpretation (general)

- Each coefficient has to be interpreted relative to the baseline
- Continuous predictor:
    - β_{1j} is the increase (or decrease) in the log-odds of $Y = j$ vs $Y = 1$ when increasing x_1 by one unit
    - $e^{\beta_{1j}}$ is the multiplicative increase (or decrease) in the odds of $Y = j$ vs $Y = 1$ when increasing x_1 by one unit
- Binary predictor:
    - β_{1j} is the log-odds of $Y = j$ vs $Y = 1$ for the group with $x_1 = 1$ compared to the group with $x_1 = 0$
    - $e^{\beta_{1j}}$ is the odds of $Y = j$ vs $Y = 1$ for the group with $x_1 = 1$ compared to the group with $x_1 = 0$

Apply to our example data

Write out the multinomial logistic regression models in terms of our example data (using “ses” and “write” as predictors and “academic” as the reference level)

- Start by defining J and its specific levels, p, and $\pi_{i1} \ldots \pi_{iJ}$
- Then write the $J - 1$ models with the formula on the “multinomial logistic regression setup” slide

Implementation in R

```
> library(foreign)
> library(nnet)
```

```
> ml$prog2 <- relevel(ml$prog, ref="academic")
> test <- multinom(prog2~ses+write, data=ml)
# weights:  15 (8 variable)
initial  value 219.722458
iter  10 value 179.982880
final  value 179.981726
converged
> summary(test)
Call:
multinom(formula = prog2 ~ ses + write, data = ml)

Coefficients:
         (Intercept)  sesmiddle    seshigh      write
general     2.852198 -0.5332810 -1.1628226 -0.0579287
vocation    5.218260  0.2913859 -0.9826649 -0.1136037

Std. Errors:
         (Intercept) sesmiddle   seshigh      write
general     1.166441 0.4437323 0.5142196 0.02141097
vocation    1.163552 0.4763739 0.5955665 0.02221996

Residual Deviance: 359.9635
AIC: 375.9635
```

Interpretation

- A one-unit increase in writing score is associated with a decrease of 0.058 in the log odds of being in general program vs academic program
- A one unit increase in writing score is associated with a 6% decrease of the odds of being in the general program compared to the academic program (how did I get this?)
- The log odds of being in the general program vs the academic program will decrease by 1.163 if moving from low to high SES
- The odds of being in the general program vs the academic program are 70% (or 0.3x) lower for high SES than low SES.

Testing coefficients

```
> z <- summary(test)$coefficients/summary(test)$standard.errors
> (p <- (1-pnorm(abs(z)))*2)
          (Intercept) sesmiddle    seshigh        write
general  0.0144766100 0.2294379 0.02373856 6.818902e-03
vocation 0.0000072993 0.5407530 0.09894976 3.176045e-07
```

Model assessment

- Assumptions
- Model fit
- Predictions

Predictions

```
> head(test$fitted.values)
   academic   general  vocation
1 0.1482764 0.3382454 0.5134781
2 0.1202017 0.1806283 0.6991700
3 0.4186747 0.2368082 0.3445171
4 0.1726885 0.3508384 0.4764731
5 0.1001231 0.1689374 0.7309395
6 0.3533566 0.2377976 0.4088458
```

```
> head(predict(test))
[1] vocation vocation academic vocation vocation vocation
Levels: academic general vocation
```

```
> confusionMatrix(predict(test),ml$prog2,mode="everything")
Confusion Matrix and Statistics

          Reference
Prediction academic general vocation
  academic       92      27       23
  general         4       7        4
  vocation        9      11       23

Overall Statistics

               Accuracy : 0.61
                 95% CI : (0.5387, 0.678)
    No Information Rate : 0.525
    P-Value [Acc > NIR] : 0.009485

                  Kappa : 0.2993

 Mcnemar's Test P-Value : 7.654e-06

Statistics by Class:

                     Class: academic Class: general Class: vocation
Sensitivity                   0.8762         0.1556          0.4600
Specificity                   0.4737         0.9484          0.8667
Pos Pred Value                0.6479         0.4667          0.5349
Neg Pred Value                0.7759         0.7946          0.8280
Precision                     0.6479         0.4667          0.5349
Recall                        0.8762         0.1556          0.4600
F1                            0.7449         0.2333          0.4946
Prevalence                    0.5250         0.2250          0.2500
```

```
> pred_df <- data.frame(ses=rep(c("low","middle","high"),each=31), write=rep(c(35:65),3))
> preds <- cbind(pred_df, predict(test, newdata=pred_df, type="probs"))
> head(preds)
  ses write  academic   general  vocation
1 low    35 0.1482764 0.3382454 0.5134781
2 low    36 0.1601567 0.3447838 0.4950595
3 low    37 0.1726885 0.3508384 0.4764731
4 low    38 0.1858692 0.3563633 0.4577674
5 low    39 0.1996911 0.3613150 0.4389939
6 low    40 0.2141410 0.3656530 0.4202060
> preds_long <- gather(preds, "level", "probability", 3:5)
> head(preds_long)
  ses write    level probability
1 low    35 academic   0.1482764
2 low    36 academic   0.1601567
3 low    37 academic   0.1726885
4 low    38 academic   0.1858692
5 low    39 academic   0.1996911
6 low    40 academic   0.2141410
> ggplot(preds_long, aes(x=write, y=probability, col=ses))+geom_line()+facet_grid(level~.)
```