Lesson 5 - Probabilistic Grammar

Erin M. Buchanan

02/06/2019

Language Topics Discussed

- Grammatical slots part 2!
- Understanding word choice based on contextual features

Synonyms

- How do we decide which word to use when we have multiple words that share the same meaning?
- I _____ to go to the mall this week.
  - Planned
  - Decided
  - Scheduled
  - am going to go
- She is very _______.
  - Pretty
  - Beautiful
  - Alluring

More examples

- May versus might: depending on certainty
- Gown versus dress: formality in social situation
- Soda versus coke versus pop: cultural factors

Causative Constructions

- Phrases like: might/have/get/cause X to do Y
- Combines conjugated 'to have' or 'to get' + direct object + main verb
- When one does not carry out an action oneself but rather has the action done by someone else
- I *had* him *paint* my house.
- Causee is the actor, does the action
- Causer is the person/thing who required the action

Causative Constructions

- Do versus let in Dutch
- Do appears to be a direct causation
  - "If the energy is put in, the result is inevitable"
  - He reminded me of my father (causation is involuntary)
- Let appears to be an indirect causation
  - Enablement and permission
  - I let him paint my house

Cause for Cause

- Verb:
  - State or action: does it apply to state verbs or action verbs
  - Transitivity: why types of verbs, transitive, intransitive or both?

Cause for Cause

- The action being caused:
  - Control: does the causee have control?
  - Volition: does the causee act willingly?
  - Affectedness: how is the causee affected?

Cause for Cause

- Related to the causer:
  - Directness: of the causer
  - Intention: accidental or intentional
  - Natural: natural activity or with effort
  - Involvement: the causer's involvement in activity

Criteria for Do Let

- Inducive: mental causer (human) to mental causee (let)
- Volitional: mental causer to non-mental causee (neither)
- Affective: non-mental causer to mental causee (do)
- Physical: non-mental causer to non-mental causee (do)

Logistic Regression

Original regression model examining a continuous outcome measure y:

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} ... + \epsilon_i$$

Logistic regression model examining a categorical outcome measure y:

$$g(y) = b_0 + b_1 x_{1i} + b_2 x_{2i} ... + \epsilon_i$$

Logistic Regression

- Main distinction is g(y): which is the logit or log odds of the outcome.
- Two options: binomial logistic regression
- More than two options: multinomial or polytomous regression
- g(y) represents the odds of one choice over another.

Logistic Regression

- Otherwise, information is the same:
  - b_0: intercept, chances of outcome when all predictors are zero
  - b_1: slope for one X variable, coefficient
  - ... etc. But we will also add better interpretation for the predictors to help understand likelihood of outcome (i.e., since the data is categorical)

Requirements for Logistic Regression

- Large enough sample size for the outcome variable
- How big?

```r
library(Rling)
data("doenLaten")
table(doenLaten$Aux)
```

```
##
## laten  doen
##   277   178
```

Running a Binary Logistic Regression

```r
#install.packages("rms") #if you have not used it before
library(rms)
#you can also use glm() for log regression but rms has coo
head(doenLaten)
```

```
##     Aux Country  Causation EPTrans EPTrans1
## 1 laten      NL   Inducive    Intr     Intr
## 2 laten      NL   Physical    Intr     Intr
## 3 laten      NL   Inducive      Tr       Tr
## 4  doen      BE  Affective    Intr     Intr
## 5 laten      NL   Inducive      Tr       Tr
## 6 laten      NL Volitional    Intr     Intr
```

Run your model!

```r
model = lrm(Aux ~ Causation + EPTrans + Country, #model fo
            data = doenLaten)
model
```

```
## Logistic Regression Model
##
##  lrm(formula = Aux ~ Causation + EPTrans + Country, data
##
##                        Model Likelihood     Discriminat
##                           Ratio Test           Indexes
##  Obs            455    LR chi2     271.35    R2       0.6
##   laten         277    d.f.             5    g        2.2
##   doen          178    Pr(> chi2) <0.0001    gr       9.9
##  max |deriv| 1e-07                            gp       0.3
##                                               Brier    0.1
##
##                        Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept              1.8631 0.3771  4.94  <0.0001
##  Causation=Inducive    -3.3725 0.3741 -9.01  <0.0001
```

Frequency

- Important for data screening

```
model$freq
```

```
## laten  doen
##   277   178
```

Overall Predictiveness

- Likelihood ratio test = χ^2 test
- Similar to the F-test in regression
- $\chi^2(5) = 271.35$, p $< .001$
- Compared to a model of no predictors, should have less error (deviance)

```r
model$stats
```

```
##                Obs        Max Deriv       Model L.R.
## 455.0000000000000    0.0000001119568 271.3508063697709
##                  P                  C              Dxy
##    0.0000000000000    0.8936539163591   0.7873078327181
##              Tau-a                 R2            Brier
##    0.3758435397202    0.6088475272089   0.1116057213797
##                 gr                 gp
##    9.9349009305359    0.3782199942877
```

Goodness of Fit

- Effect size of how well the model fit the data
- R^2 - much like regression, albeit more difficult to interpret.
  - In the output that's Nagelkerke's pseudo-R^2

```
model$stats
```

```
##                 Obs        Max Deriv      Model L.R.
## 455.0000000000000   0.0000001119568 271.3508063697709
##                   P                C             Dxy
##   0.0000000000000   0.8936539163591   0.7873078327181
##               Tau-a               R2           Brier
##   0.3758435397202   0.6088475272089   0.1116057213797
##                  gr               gp
##   9.9349009305359   0.3782199942877
```

Goodness of Fit

- Concordance index C
  - For each Y_i, a probability of the outcome is created
  - C is the number of times that the probability of the outcome matches the actual outcome
- Interpretation:
  - $<$ .5: no discrimination
  - .7 $<=$ C $<$ .8: acceptable
  - .8 $<=$ C $<$ .9: excellent
  - $<=$ .9 outstanding

Coefficients

- Under Coef, these are represented as log odds
- Odds ratios are centered around one (like 6 to 1, 4 to 1)
- Log odds ratio are centered around zero
  - Positive numbers indicate a higher probability for the coded group
  - Negative numbers indicate a higher probability for the comparison group
- What is coded versus comparison?

```
levels(doenLaten$Aux)
```

```
## [1] "laten" "doen"
```

Coefficients

```
#model
 #                            Coef    S.E.   Wald Z Pr(>|Z|)
 # Intercept                   1.8631 0.3771  4.94  <0.0001
 # Causation=Inducive         -3.3725 0.3741 -9.01  <0.0001
 # Causation=Physical          0.4661 0.6275  0.74  0.4576
 # Causation=Volitional       -3.7373 0.4278 -8.74  <0.0001
 # EPTrans=Tr                 -1.2952 0.3394 -3.82  0.0001
 # Country=BE                  0.7085 0.2841  2.49  0.0126
```

Coefficients

```
# Causation=Inducive   -3.3725 0.3741 -9.01  <0.0001
# Causation=Physical    0.4661 0.6275  0.74  0.4576
# Causation=Volitional -3.7373 0.4278 -8.74  <0.0001

table(doenLaten$Aux, doenLaten$Causation)
```

```
##
##          Affective Inducive Physical Volitional
##   laten         15      160        4         98
##   doen          75       25       59         19
```

Coefficients

```
 # EPTrans=Tr                 -1.2952 0.3394 -3.82  0.0001
 # Country=BE                  0.7085 0.2841  2.49  0.0126

table(doenLaten$Aux, doenLaten$EPTrans)
```

```
##
##          Intr  Tr
##   laten   137 140
##   doen    144  34
```

```
table(doenLaten$Aux, doenLaten$Country)
```

```
##
##           NL  BE
##   laten  162 115
##   doen    71 107
```

Interactions

```
model1 = glm(Aux ~ Causation + EPTrans + Country, #model fo
             family = binomial,
             data = doenLaten)
model2 = glm(Aux ~ Causation + EPTrans*Country, #model form
             family = binomial,
             data = doenLaten)
anova(model1, model2, test = "Chisq")
```

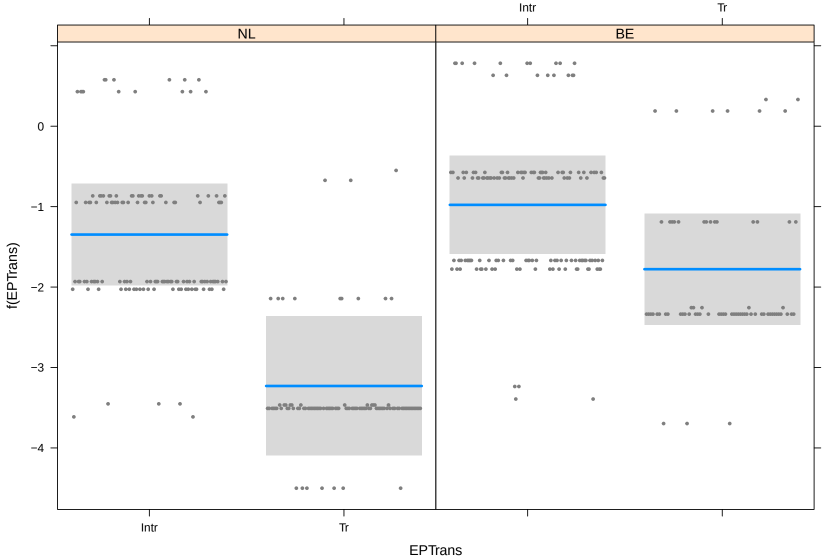
```
## Analysis of Deviance Table
##
## Model 1: Aux ~ Causation + EPTrans + Country
## Model 2: Aux ~ Causation + EPTrans * Country
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       449     337.70
## 2       448     334.58  1   3.1151  0.07757 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Interactions

```
#summary(model2)
# Coefficients:
#                        Estimate Std. Error z value
# (Intercept)              2.0991     0.4079   5.146
# CausationInducive       -3.4463     0.3892  -8.854 < 0.00000
# CausationPhysical        0.3898     0.6336   0.615
# CausationVolitional     -3.7795     0.4364  -8.661 < 0.00000
# EPTransTr               -1.8825     0.4919  -3.827
# CountryBE                0.3693     0.3416   1.081
# EPTransTr:CountryBE      1.0827     0.6215   1.742
```

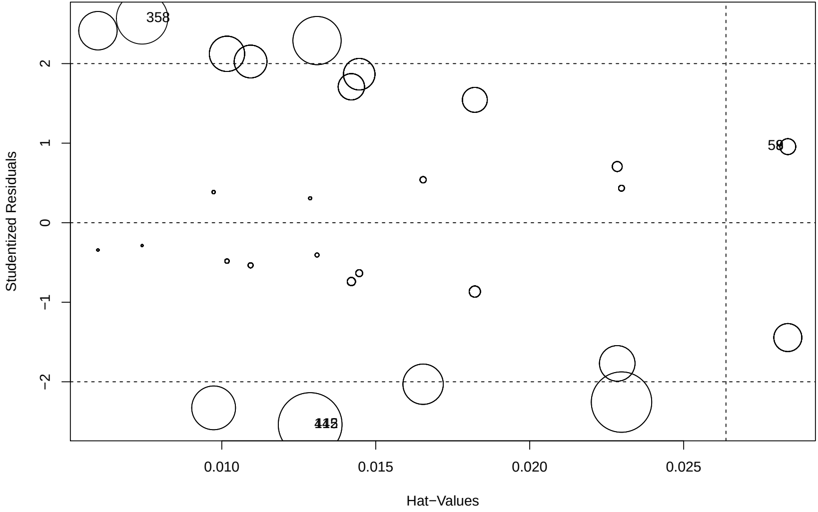
Interactions

```
library(visreg)
visreg(model2, "EPTrans", by = "Country")
```



Outliers

```r
library(car)
influencePlot(model1)
```




```
##              StudRes           Hat         CookD
```

Assumptions

- Observations are independent
- No multicollinearity (remember VIF $>$ 5 or 10 is bad)
- Overprediction (complete/quasi complete separation)

```
rms::vif(model) #use the rms:: to distinguish between car:
```

```
##   Causation=Inducive   Causation=Physical Causation=Vol
##             1.699064             1.356411            1.
##           EPTrans=Tr           Country=BE
##             1.270669             1.017354
```

Summary

- We can model word choice with various predictors by using logistic regression with two outcomes.
  - You can extend that to multinomial logistic regression with `mlogit`
  - https://www.youtube.com/watch?v=c78eMWw43I0 is a tutorial from Dr. B (if you are interested for your final project)
- You learned how to run and use a logistic regression, along with assumptions checks and understanding the output