Part 5
Introduction to logistic regression

Natalia Levshina © 2015

Part of the course taught at the University of Mainz, Germany
26-28 May 2015

Outline

1. Data: Dutch causative constructions
2. Binomial logistic regression
   - principles
   - main functions
   - variable selection
   - interactions
   - how many observations are needed?

A case study

- Causative verbs doen "do, make" and laten "let"
- Semantics: doen expresses more direct causation than laten
- Syntax: doen is used more often with intransitive verbs
- Geographic variation: causative doen occurs more frequently in Belgian Dutch

(1)	*Hij*	*deed*	*me*	*denken*	*aan*	*mijn*	*vader.*
	He	did	me	think	at	my	father
	"He reminded me of my father."						
(2)	*Ik*	*liet*	*hem*	*mijn*	*huis*	*schilderen.*	
	I	let	him	my	house	paint	
	"I had him paint my house."						

Data format for logistic regression

```
> library(Rling)
> data(doenLaten)
> 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
```

Outline

1. Data: Dutch causative constructions
2. Binomial logistic regression
    - principles
    - main functions
    - variable selection
    - interactions
    - how many observations are needed?

Logistic regression

- Models the relationship between a categorical response (e.g. doen or laten, active or passive voice, going to or gonna) and one or more predictors (e.g. direct or indirect causation, spoken or written data, the country, formal or informal speech…)

  - Two outcomes: binomial (dichotomous)

  - Three and more: multinomial (polytomous)

Outline

1. Multifactorial grammar
2. Binomial logistic regression
   - principles
   - main functions
   - variable selection
   - interactions
   - how many observations are needed?

Two most useful functions

- `glm()` from the basic distribution

For example:

```
> your.glm <- glm(Outcome ~ PredictorX +
PredictorY + …, family = binomial, data =
yourData)
> summary(your.glm)
```

- `lrm()` from package `rms` by Frank Harrell

For example:

```
> your.lrm <- lrm(Outcome ~ PredictorX +
PredictorY + …, data = yourData)
> your.lrm
```

A lrm model

```
> library(rms) #install it first: Main Menu: Packages > Install package
> m.lrm <- lrm(Aux ~ Causation + EPTrans + Country, data = doenLaten)
> m.lrm
…
```

Interpreting the output

(1)

```
Logistic Regression Model
lrm(formula = Aux ~ Causation + EPTrans
+ Country, data = doenLaten)
```

(2)

```
Obs            455 # total number of obs.
laten          277   and each outcome.
doen           178
## The first level (laten) is the
reference level!!!
```

Interpreting the output (cont.)

```
(3)
Model Likelihood
Ratio Test
LR chi2      271.35
d.f.              5
Pr(> chi2) <0.0001  #overall model significance
```

The null hypothesis of the test is that the **deviance** (i.e. unexplained variation in logistic regression) of the current model does not differ from the deviance of a model without any predictors.

Interpreting the output (cont.)

```
(4)
Discrimination
Indexes
R2        0.609 #pseudo-R2: from 0 (no
predictive power) to 1 (perfect
prediction)
…
(5)
Rank Discrim.
Indexes
C        0.894 #Concordance index C
…
```

More on C index

- If you take all possible pairs that contain a sentence with *doen* and a sentence with *laten*, and try all combinations, the statistic C will be the proportion of the times when the model predicts a higher probability of *doen* for the sentence with *doen* , and a higher probability of *laten* for the sentence with *laten.*
- Rule of thumb:

$C = 0.5$	no discrimination
$0.7 \leq C < 0.8$	acceptable discrimination
$0.8 \leq C < 0.9$	excellent discrimination
$C \geq 0.9$	outstanding discrimination

Table of coefficients

	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

Interpretation of coefficients

- The coefficients are log odds ratios
- A positive coefficient means that the feature increases the chances of doen in comparison with laten, other things being equal (laten is the reference level!).
- A negative coefficient shows that the feature increases the chances of laten in comparison with doen, other things being equal.
- Dummy coding for categorical variables: each level is compared with the reference level (Causation = Affective, EPTrans = Intr, Country = NL)

Outline

1. Multifactorial grammar
2. Binomial logistic regression

   - principles

   - main functions

   - variable selection

   - interactions

   - how many observations are needed?

Variable selection: strategies

- Usually, we strive for a parsimonious model, i.e. a model where every predictor is useful, there is no redundancy
- Two popular strategies:
  - Theory-driven (all variables of interest)
  - Stepwise

Stepwise

- Forward (adding predictors one by one until there is no more improvement)
- Backward (removing predictors one by one until the model becomes significantly worse)
- Bidirectional (a combination of the two)

- The main criterion: AIC (Akaike Information Criterion), a measure of model quality with regard to the number of predictors. A trade-off between model complexity and goodness of fit (cf. R2, C index). The smaller AIC for the same data, the better.

Stepwise selection

```
> m0.glm <- glm(Aux ~ 1, data =
doenLaten, family = binomial) # model
with intercept only
> m.fw <- step(m0.glm, direction =
"forward", scope = ~ Causation + EPTrans
+ Country) # forward selection
...
> m.glm <- glm(Aux ~ Causation + EPTrans
+ Country, data = doenLaten, family =
binomial) # a full glm model
> m.bw <- step(m.glm, direction =
"backward") # backward elimination
…
```

Stepwise selection (cont.)

```
> m.both <- step(m.glm) # bidirectional
by default
…
```

Outline

1. Multifactorial grammar
2. Binomial logistic regression
   - principles
   - main functions
   - variable selection
   - interactions
   - how many observations are needed?

Testing interactions

- Interaction of two or more predictors means that their effect is not additive
- Consider Belgian fries and Belgian chocolate: both delicious, but not so much if you try to eat them together.
- To test interactions, you can use anova():

An example

```
> m.glm.int <- glm(Aux ~ Causation + EPTrans*Country, data = doenLaten, family = binomial)
> anova(m.glm, m.glm.int, 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 ‘ ’ 1
```

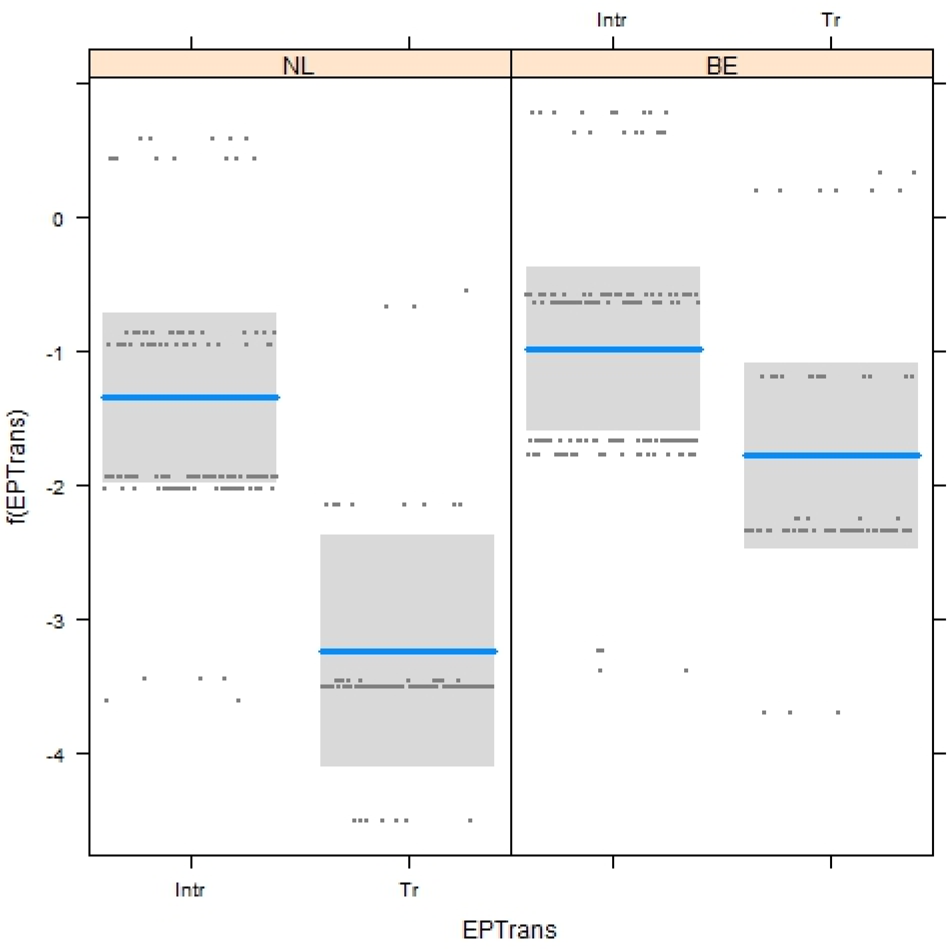
An example

```
> m.glm.int <- glm(Aux ~ Causation +
EPTrans*Country, data = doenLaten, family  =
binomial)
> anova(m.glm, m.glm.int, 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 ‘ ’ 1
```

Only marginally significant!

Interpreting an interaction

```
> library(visreg) # install the package first!
> visreg(m.glm.int, "EPTrans", by = "Country")
```



NL
BE
Intr
Tr
0
-1
-2
-3
-4
f(EPTrans)
Intr
Tr
EPTrans

Outline

1. Multifactorial grammar
2. Binomial logistic regression

    - principles

    - main functions

    - variable selection

    - interactions

    - how many observations are needed?

How many observations are needed?

- If there are too few observations, the model will be overfitted. This means that it will be useless when it is applied to new data.
- Rule of thumb: not less than 10 obs. with the LESS frequent outcome per parameter in the model (see the regression coefficients).

```
> summary(doenLaten$Aux)
laten   doen
 277    178
```

- e.g. 6 parameters x 10 = 60
- 60 < 178, seems OK!