

Multinomiale Logistische Regression in R

Benjamin Schlegel

16. Mai 2016

Eine multinomiale Regression kann in R mit der Funktion `multinom()` aus der Bibliothek `nnet` geschätzt werden. Doch bevor ein Modell geschätzt werden kann, müssen die Daten vorbereitet werden und eine Referenzkategorie definiert werden. Es soll überprüft werden, ob die Anzahl Autos einen Indikator ist, welche Ethnie eine Person hat. Der London Data Datensatz kann unter [Data](#) heruntergeladen werden.

```
data = read.csv("london_survey_2010.csv")

Q3 = as.character(data$Q3)
data$ethnie = ifelse(grepl("WHITE",Q3),"WHITE",
                    ifelse(grepl("ASIAN",Q3),"ASIAN",
                            ifelse(grepl("BLACK",Q3),"BLACK",
                                    ifelse(grepl("MIXED",Q3),"MIXED",NA))))

data$ethnie = factor(data$ethnie)
data$car = data$Q62
```

Der obige Code liest den Datensatz ein und fasst die verschiedenen Ethnien zu vier Ethnien zusammen. Die Referenzkategorie kann mit der Funktion `relevel()` definiert werden. WHITE wird als Referenzkategorie definiert, da es viele Fälle hat und vermutlich am weitesten vom Rest entfernt ist.

```
data$ethnie = relevel(data$ethnie, ref="WHITE")
```

Nun sind alle Vorbereitungen abgeschlossen und das Modell kann geschätzt werden.

```
library(nnet)
model = multinom(ethnie ~ car, data=data)
summary(model)
```

Call:

```
multinom(formula = ethnie ~ car, data = data)
```

Coefficients:

	(Intercept)	car2	cars/light vans	car3+	cars/light vans	carNone
ASIAN	-1.241919		-0.3743366		0.7026788	-0.4079142
BLACK	-1.794023		-0.8870796		-9.6106501	0.3184763
MIXED	-3.080670		-0.8048357		1.0015090	0.1777227

Std. Errors:

	(Intercept)	car2	cars/light vans	car3+	cars/light vans	carNone
ASIAN	0.09881256		0.2260396		0.3505285	0.1693659
BLACK	0.12388127		0.3495763		61.1475242	0.1781259

MIXED 0.22322625

0.6246454

0.6517021 0.3293187

Residual Deviance: 2534.756

AIC: 2558.756

Der Output des Modells bringt jedoch noch nichts, da die Richtung der Koeffizienten beim multinomialen Modell keinen Hinweis auf die Richtung des Effekts gibt. Am einfachsten ist es, vorausgesagte Wahrscheinlichkeiten und diskrete Änderungen zu berechnen. Dazu kann die Bibliothek [glm.predict](#) verwendet werden.

```
dc = predicts(model,"F", position = 1)
```

```
dc
```

	vall_mean	vall_lower	vall_upper	val2_mean	val2_lower	val2_upper	dc_mean
1	0.66480214	6.340740e-01	0.69672421	0.77011090	7.073228e-01	0.82333901	-0.105308755
	-0.168549944	-0.032806816	1	car or light van	2 cars/light vans	WHITE	
2	0.19284319	1.635913e-01	0.22419834	0.15568113	1.089938e-01	0.21399887	0.037162062
	-0.026277492	0.093203228	1	car or light van	2 cars/light vans	ASIAN	
3	0.11135945	8.932386e-02	0.13707471	0.05546606	2.880155e-02	0.09682249	0.055893391
	0.009944955	0.093739422	1	car or light van	2 cars/light vans	BLACK	
4	0.03099522	1.982287e-02	0.04611341	0.01874192	5.054101e-03	0.04724902	0.012253302
	-0.018390666	0.034157628	1	car or light van	2 cars/light vans	MIXED	
5	0.66485759	6.308629e-01	0.69446062	0.33170952	8.319508e-46	0.69613358	0.333148069
	-0.026094801	0.687818692	1	car or light van	3+ cars/light vans	WHITE	
6	0.19235360	1.639421e-01	0.22297750	0.19864139	6.426502e-46	0.47118646	-0.006287795
	-0.284959332	0.218449682	1	car or light van	3+ cars/light vans	ASIAN	
7	0.11145916	8.910456e-02	0.13789452	0.42038715	5.390543e-59	1.00000000	-0.308927988
	-0.907150024	0.134830604	1	car or light van	3+ cars/light vans	BLACK	
8	0.03132965	1.984111e-02	0.04632367	0.04926194	9.625615e-47	0.18506932	-0.017932286
	-0.151445971	0.042469291	1	car or light van	3+ cars/light vans	MIXED	
9	0.66507106	6.332480e-01	0.69742878	0.67634403	6.373734e-01	0.71371323	-0.011272973
	-0.062551449	0.040825919	1	car or light van		None WHITE	
10	0.19247841	1.626337e-01	0.22352551	0.13012617	1.006508e-01	0.16608856	0.062352241
	0.014655704	0.104437094	1	car or light van		None ASIAN	
11	0.11137474	8.933627e-02	0.13727125	0.15549655	1.250712e-01	0.18665898	-0.044121816
	-0.084553457	-0.005390108	1	car or light van		None BLACK	
12	0.03107580	1.973303e-02	0.04716460	0.03803325	2.322049e-02	0.05787455	-0.006957451
	-0.029743470	0.015067765	1	car or light van		None MIXED	
13	0.77012658	7.105816e-01	0.82109678	0.33047487	1.048114e-48	0.69753371	0.439651706
	0.049275130	0.811099392	2	cars/light vans	3+ cars/light vans	WHITE	
14	0.15550726	1.090630e-01	0.20864517	0.19478384	7.021075e-49	0.48114435	-0.039276582
	-0.330217191	0.198550213	2	cars/light vans	3+ cars/light vans	ASIAN	
15	0.05634516	2.878989e-02	0.09832707	0.42835632	5.140888e-57	1.00000000	-0.372011156
	-0.966024988	0.093482419	2	cars/light vans	3+ cars/light vans	BLACK	
16	0.01802100	4.738508e-03	0.04531185	0.04638497	2.358185e-49	0.17968855	-0.028363968
	-0.164343249	0.037750776	2	cars/light vans	3+ cars/light vans	MIXED	
17	0.76969432	7.075275e-01	0.82206545	0.67552171	6.371199e-01	0.71267436	0.094172613
	0.022828203	0.163287390	2	cars/light vans		None WHITE	
18	0.15543525	1.070275e-01	0.20497428	0.13154406	1.032694e-01	0.16306163	0.023891195
	-0.034068377	0.082428402	2	cars/light vans		None ASIAN	
19	0.05631627	2.729855e-02	0.09548009	0.15476924	1.215174e-01	0.18861677	-0.098452968

```
-0.143296287 -0.046905270 2 cars/light vans None BLACK
20 0.01855415 5.160892e-03 0.04908724 0.03816499 2.351929e-02 0.05802048 -0.019610839
-0.045552843 0.014590945 2 cars/light vans None MIXED
21 0.31648549 1.832143e-51 0.69659445 0.67541612 6.365706e-01 0.71052080 -0.358930629
-0.703906178 0.023696500 3+ cars/light vans None WHITE
22 0.18351632 1.100744e-51 0.45212966 0.13131460 1.027665e-01 0.16408443 0.052201719
-0.156822146 0.327383712 3+ cars/light vans None ASIAN
23 0.45491113 2.499555e-57 1.00000000 0.15495677 1.250609e-01 0.19104058 0.299954355
-0.182502024 0.867128525 3+ cars/light vans None BLACK
24 0.04508706 4.490685e-52 0.17085489 0.03831251 2.246111e-02 0.05905635 0.006774555
-0.054724747 0.139231954 3+ cars/light vans None MIXED
```

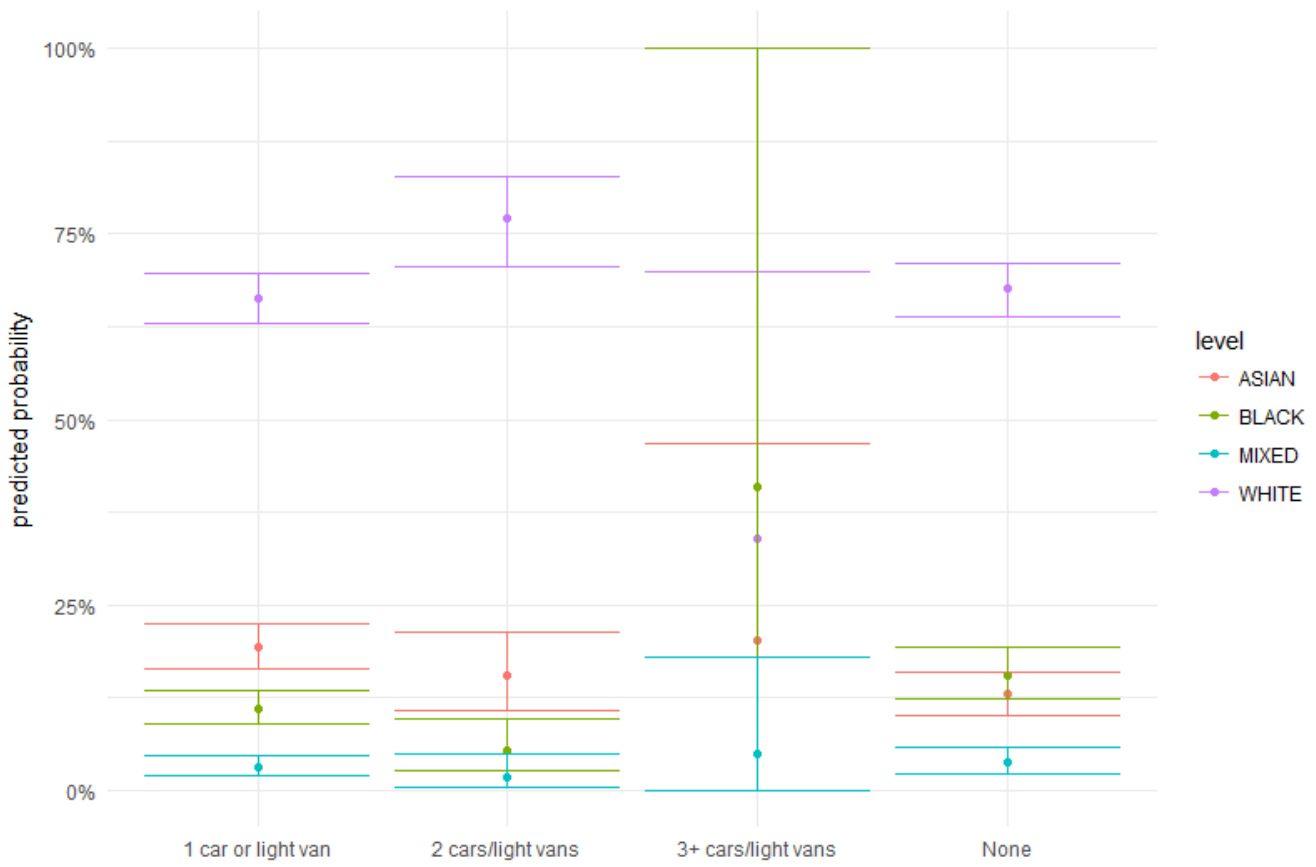
Aus dem Output kann beispielsweise abgelesen werden, dass bei zwei Autos es eher ein Weisser ist als bei nur einem oder drei Autos. Hingegen macht der Unterschied zwischen einem und zwei Autos keinen Unterschied auf die Wahrscheinlichkeit, dass es sich um einen Asiaten handelt.

Als letzter Schritt werden die vorausgesagten Wahrscheinlichkeiten noch geplottet. Dazu wird einfach position weggelassen.

```
library(ggplot2)
library(scales)
```

```
pred.prop = predicts(model, "F")
```

```
ggplot(pred.prop, aes(x = car, color = level)) + geom_point(aes(y=mean)) +
  geom_errorbar(aes(ymin = lower, ymax = upper)) + theme_minimal() +
  ylab("predicted probability") + xlab("cars") +
  scale_y_continuous(labels = percent)
```



Will man Modelle vergleichen, kommt beispielsweise der BIC in Frage. Je kleiner der BIC ist, desto besser das Modell. Der BIC kann mit der Funktion `BIC()` berechnet werden. Bei genesteten Modellen kann auch ein `wald.test` gerechnet werden, um herauszufinden, ob die zusätzlichen Variablen einen Mehrwert bringen.