Simulation in R für eine logistische Regression

Benjamin Schlegel

14. November 2016

In diesem Artikel wird gezeigt, wie man in R simulieren kann, ohne die Bibliothek glm.predict zu verwenden. Der Artikel geht davon aus, dass man schon weiss, was eine logistische Regression ist und wie man sie schätzt. Die Simulation ist eine Möglichkeit, um von vorausgesagten Wahrscheinlichkeiten und diskreten Änderungen Konfidenzintervalle zu berechnen. Die Simulation beruht darauf, dass die Fehler asymtotisch normalverteilt sind. Asymptotisch bedeutet, dass es unendlich viele Fälle gibt. In der Realität ist dies natürlich nie der Fall. Bei steigender Fallzahl gleichen sich die Fehler jedoch der Normalverteilung an, weshalb die Simulation bei einer hohen Fallzahl verwendet werden kann.

Für das Beispiel werden zuerst die Daten eingelesen, recodiert und anschliessend ein logistisches Modell gerechnet. Das Modell schätzt den Einfluss von Europaskeptizismus und dem Alter auf die Wahrscheinlichkeit die Labour Partei in Grossbritannien zu wählen.

```
data = effects::BEPS
head(data)
data$labour = car::recode(data$vote,"'Labour'=1;else=0")
model1 = glm(labour ~ Europe + age,data=data,family=binomial)
summary(model1)
```

```
Call:
glm(formula = labour ~ Europe + age, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6359  -1.0601  -0.7896   1.1047   1.6623

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.439146   0.217837   6.607 3.93e-11 ***
Europe      -0.173531   0.016693 -10.395  < 2e-16 ***
age         -0.007241   0.003417  -2.119   0.0341 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2109.4  on 1524  degrees of freedom
Residual deviance: 1986.2  on 1522  degrees of freedom
AIC: 1992.2

Number of Fisher Scoring iterations: 4
```

Nun interessiert uns wie Wahrscheinlichkeit einer Person mit durchschnittlichem Alter und Euroskeptizismus die Labour Partei wählt. Zuerst müssen wir einen Vektor mit den Werten definieren für jedes β. β_0 ist immer 1,

β_1 ist in diesem Fall 1 (wenig Euroskeptizismus) und β_2 54.18 (durchschnittliches Alter).

```
values = c(1,1,54.18)
```

Anschliessend wir die eigentliche Simulation durchgeführt indem 1000 (oder mehr) zufällige Betas aus der multivariaten Normalverteilung gezogen werden. Dazu kann die Funktion mvrnorm aus der Bibliothek MASS verwendet werden. Der erste Parameter gibt die Anzahl Fälle an, der zweite die geschätzten Betas des Modells und als drittes Argument die Varianz-Covarianz-Matrix.

```
library(MASS)
beta = mvrnorm(1000,coef(model1),vcov(model1))
```

Im nächsten Schritt können wir für die 1000 Fälle die vorausgesagte Wahrscheinlichkeit mit der inversen Linkfunktion berechnen.

$$\pi_i = \frac{1}{1 + e^{-x_i^T \beta}} = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}$$

```
pi = c()
for(i in 1:1000){
  xBeta = values %*% beta[i,]
  pi[i] = exp(xBeta) / (1 + exp(xBeta))
}
```

Nun können der Durchschnitt und das Konfidenzintervall dieser 1000 Fälle berechnet werden.

```
mean(pi)
```

```
[1] 0.7054789
```

```
quantile(pi,probs = c(.025,.975))
```

```
     2.5%     97.5% 
0.6601499 0.7489786 
```

Mit dem gleichen Verfahren können auch diskrete Änderungen berechnet werden, um herauszufinden, ob der Unterschied zwischen zwei vorausgesagten Wahrscheinlichkeiten signifikant ist. Dazu werden 1000 Fälle für Fall 1 und Fall 2 berechnet. Anschliessend wird die Differenz dieser 1000 Fälle genommen. Von diesem 1000 Unterschieden können wiederum Durchschnitt und Konfidenzintervall berechnet werden.

```
values.1 = c(1,1,54.18)
values.2 = c(1,11,54.18)
pi.1 = pi.2 = c()

for(i in 1:1000){
```

```
  xBeta.1 = values.1 %*% beta[i,]
  pi.1[i] = exp(xBeta.1) / (1 + exp(xBeta.1))

  xBeta.2 = values.2 %*% beta[i,]
  pi.2[i] = exp(xBeta.2) / (1 + exp(xBeta.2))
}

discrete.change = pi.1 - pi.2
mean(discrete.change)
```

```
[1] 0.4044262
```

```
quantile(discrete.change,probs = c(.025,.975))
```

```
     2.5%     97.5%
0.3327310 0.4758304
```