Regresja poissonowska

Piotr J. Sobczyk

2 lutego 2017

Powtórka porządkująca

Na poprzednich zajęciach zajmowaliśmy się regresją logistyczną, to znaczy staraliśmy się rozdzielić obserwacje należące do dwóch kategorii, za pomocą zmiennych objaśniających.

Powtórzmy sobie najważniejsze elementy, tego jak modelujemy, czyli opisujemy na matematyczną modłę, dane w regresji logistycznej.

$$y \sim B(1, p),$$

gdzie p jest rzecz jasna wartością oczekiwaną y, i zadane jest przez funkcję od kombinacji liniowej zmiennych objaśniających

$$p = E(y|X) = \frac{exp(X\beta)}{1 + exp(X\beta)}$$

lub równoważnie

$$logit(p) = logit(E(y|X)) = x^T\beta$$

Dla regresji liniowej mamy,

$$y \sim N(E(y), \sigma^2),$$

gdzie $E(y) = x^T\beta$.

Zauważmy zatem, że nasz model zadany jest przez dwa „parametry". Pierwszy to rozkład z jakiego pochodzi zmienna objaśniana. Drugi to funkcja jaka łączy wartość oczekiwaną w rozkładzie z liniową kombinacją zmiennych objaśniających.

Te dwa elementy, rodzina rozkładów i funkcja łącząca (link function) zadają nam uogólniony model liniowy. Praktycznie jedynym ograniczeniem jakie mamy przy tworzeniu w ten sposób modeli jest to, czy potrafimy policzyć estymator parametru β.

Wstęp motywujący

W roku 1898 Ladislaus Bortkiewicz, który był rzecz jasna rosyjskim ekonomistą, opublikował w Lipsku książkę, w której przedstawił dane dotyczące liczby żołnierzy armii pruskiej, którzy umarli od kopnięcia konia na przestrzeni 20 lat. Dane można pobrać tu.

Pytanie jak modelować tego typu dane:

```
library(dplyr)
library(tidyr)
library(stringr)
bortkiewicz=read.table("bortkiewicz.csv", sep="\t")
knitr::kable(head(bortkiewicz))
```

	X1875	X1876	X1877	X1878	X1879	X1880	X1881	X1882	X1883	X1884	X1885	X1886	X1887	X
G	0	2	2	1	0	0	1	1	0	3	0	2	1	
I	0	0	0	2	0	3	0	2	0	0	0	1	1	
II	0	0	0	2	0	2	0	0	1	1	0	0	2	
III	0	0	0	1	1	1	2	0	2	0	0	0	1	

	X1875	X1876	X1877	X1878	X1879	X1880	X1881	X1882	X1883	X1884	X1885	X1886	X1887	X
IV	0	1	0	1	1	1	1	0	0	0	0	1	0	
V	0	0	0	0	2	1	0	0	1	0	0	1	0	

```
table(unlist(bortkiewicz))
```

```
##
##   0   1   2   3   4
## 144  91  32  11   2
```

```
bortkiewicz$corp=rownames(bortkiewicz)
gather(bortkiewicz, year, value, -corp) %>%
  mutate(year=as.numeric(str_extract(year, pattern="[0-9]+"))) -> bortkiewicz
summary(m1 <- glm(value ~ year + corp, family="poisson", data=bortkiewicz))
```

```
##
## Call:
## glm(formula = value ~ year + corp, family = "poisson", data = bortkiewicz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6887  -1.1077  -0.8035   0.5348   2.0810
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.559e+01  2.343e+01  -1.519   0.1288
## year         1.876e-02  1.243e-02   1.510   0.1312
## corpI        3.850e-09  3.535e-01   0.000   1.0000
## corpII      -2.877e-01  3.819e-01  -0.753   0.4512
## corpIII     -2.877e-01  3.819e-01  -0.753   0.4512
## corpIV      -6.931e-01  4.330e-01  -1.601   0.1094
## corpIX      -2.076e-01  3.734e-01  -0.556   0.5781
## corpV       -3.747e-01  3.917e-01  -0.957   0.3387
## corpVI       6.062e-02  3.483e-01   0.174   0.8618
## corpVII     -2.877e-01  3.819e-01  -0.753   0.4512
## corpVIII    -8.267e-01  4.532e-01  -1.824   0.0681 .
## corpX       -6.454e-02  3.594e-01  -0.180   0.8575
## corpXI       4.463e-01  3.202e-01   1.394   0.1633
## corpXIV      4.055e-01  3.227e-01   1.256   0.2090
## corpXV      -6.931e-01  4.330e-01  -1.601   0.1094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 323.23  on 279  degrees of freedom
## Residual deviance: 294.81  on 265  degrees of freedom
## AIC: 629.89
##
## Number of Fisher Scoring iterations: 5
```

Czy model jest sensowny? Można policzyć analogon statystyki F dla OLS. Porównujemy skonstruowany model z *saturated model* (model pełny, nasycony?).

Figure 1: Ilustracja problemu badawczego

```r
1-pchisq(summary(m1)$deviance, summary(m1)$df.residual)
```

```
## [1] 0.1006652
```

Więcej informacji na temat modeli zliczeniowych możecie znaleźć na kursie Analiza danych jakościowych. Możemy porównywać modele między sobą:

```r
anova(glm(value ~ 1, family="poisson", data=bortkiewicz), m1)
```

```
## Analysis of Deviance Table
##
## Model 1: value ~ 1
## Model 2: value ~ year + corp
##   Resid. Df Resid. Dev Df Deviance
## 1       279     323.23
## 2       265     294.81 14   28.423
```

```r
1-pchisq(28.4, 14)
```

```
## [1] 0.01258477
```

Jaki model dostaniemy stosując regreję regularyzowaną (BIC)

```r
step(m1, k = log(nrow(bortkiewicz)))
```

```
## Start:  AIC=684.41
## value ~ year + corp
##
##        Df Deviance    AIC
## - corp 13   320.94 637.29
## - year  1   297.09 681.06
## <none>      294.81 684.41
##
## Step:  AIC=637.29
## value ~ year
##
##        Df Deviance    AIC
## - year  1   323.23 633.94
## <none>      320.94 637.29
##
## Step:  AIC=633.94
## value ~ 1
```

```
##
## Call:  glm(formula = value ~ 1, family = "poisson", data = bortkiewicz)
##
## Coefficients:
## (Intercept)
##     -0.3567
##
## Degrees of Freedom: 279 Total (i.e. Null);  279 Residual
## Null Deviance:       323.2
## Residual Deviance: 323.2     AIC: 630.3
```

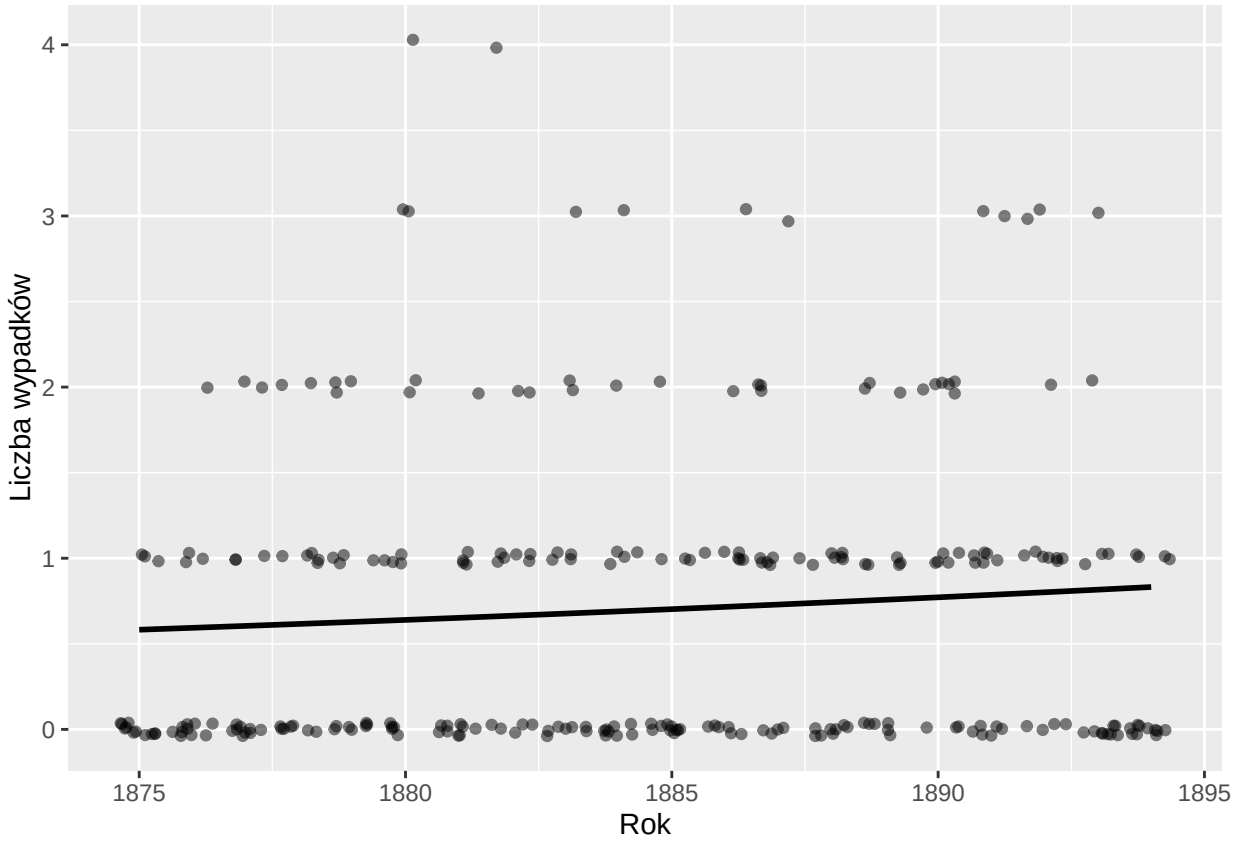
```r
m2 <- update(m1, . ~ . - corp)
## test model differences with chi square test
anova(m2, m1, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: value ~ year
## Model 2: value ~ year + corp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       278     320.94
## 2       265     294.81 13   26.137   0.0163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
1-pchisq(summary(m2)$deviance, summary(m2)$df.residual)
```

```
## [1] 0.038955
```

```
library(ggplot2)
bortkiewicz$valuehat <- predict(m2, type="response")
ggplot(bortkiewicz, aes(x = year, y = valuehat)) +
  geom_point(aes(y = value), alpha=.5, position=position_jitter(h=.1)) +
  geom_line(size = 1) +
  labs(x = "Rok", y = "Liczba wypadków")
```



Trochę teorii

Zmienna wynikowa y pochodzi z rozkładu Poissona

$$y \sim Poiss(\mu),$$

gdzie $E(y) = \mu$. Przypomnijmy co to oznacza:

1. y przyjmuje wartości całkowite nieujemne
2. $P(y = k) = \frac{\exp(\mu) \cdot \mu^y}{y!}$

μ zależy od zmiennych objaśniających przez funkcję łączącą (link function). Dla regresji poissonowskiej stosujemy dwie

- funkcję identycznościową $\mu = \beta^T x$. Uwaga! Możemy dostać $\mu < 0$
- logarytm (domyślna w R) $\log(\mu) = \beta^T x$

Interpretacja parametrów:

- $\exp(\beta_0)$ - efekt na średniej y gdy $x_{-0} = 0$
- $\exp(\beta_i)$ - zwiększenie się x_i o jeden, powoduje wzrost wartości oczekiwanej y przez przemnożenie przez $\exp(\beta_i)$. To jaki efekt ma b_i zależy od znaku.

Predykcja

Predykcje w regresji poissona wykonujemy w ten sam sposób jak dla OLS i regresji logistycznej, bierzemy wartość oczekiwaną.

```
predict(m1, newdata = bortkiewicz[1:10,])
```

```
##          1          2          3          4          5          6
## -0.4072510 -0.4072510 -0.6949330 -0.6949330 -1.1003981 -0.7819444
##          7          8          9         10
## -0.3466263 -0.6949330 -1.2339295 -0.6148903
```

```
exp(predict(m1, newdata = bortkiewicz[1:10,]))
```

```
##         1         2         3         4         5         6         7
## 0.6654772 0.6654772 0.4991079 0.4991079 0.3327386 0.4575156 0.7070695
##         8         9        10
## 0.4991079 0.2911463 0.5407002
```

```
predict(m1, newdata = bortkiewicz[1:10,], "response")
```

```
##         1         2         3         4         5         6         7
## 0.6654772 0.6654772 0.4991079 0.4991079 0.3327386 0.4575156 0.7070695
##         8         9        10
## 0.4991079 0.2911463 0.5407002
```

Ale przecież wartości mają być całkowite! Napotykamy na ten sam problem co przy regresji logistycznej. Musimy zdyskretyzować wyniki.

Rozszerzenia

Zauważmy, że jest istnieje kilka rozkładów, które nadają się do modelowania danych zliczeniowych. Z związku z tym mamy kilka rodzajów regresji:

- regresja ujemna dwumianowa. Stosujemy ją jeśli dane mają większą dyspersję niż wynikałoby to z rozkładu Poissona, w którym wariancja jest równa średniej
- regresja **zero-inflated negative binomial** lub **zero-inflated poissonn**. Używamy ich wtedy gdy rozkład ma nadreprezentację zer
- można zastosować zwykłą regresję (OLS), ale trzeba się liczyć z potrzebą transformacji danych i problemami z ciągłymi wartościami wyników (a także ujemnymi predykcjami). To robiliśmy w sprawozdaniu 2.

Jak porównać tak różne modele między sobą?

W ramach ćwiczenia polecam przejść przez analizę danych dotyczącą (cech samic krabów wpływających na liczbę będących w jej okolicy samców)[https://onlinecourses.science.psu.edu/stat504/node/169].

Inne ciekawe dane: liczba prac naukowych z biochemii http://www3.nd.edu/~rwilliam/statafiles/couart4.dta

Żródła:

1. Skrypt **Analiza danych jakościowych** dr. Andrzeja Dąbrowskiego
2. http://www.ats.ucla.edu/stat/r/dae/poissonreg.htm
3. https://onlinecourses.science.psu.edu/stat504/node/168
4. http://data.princeton.edu/wws509/notes/c4.pdf