Wykorzystanie modelu ograniczonego Poissona w badaniu wypadków drogowych w Wielkiej Brytanii

07.01.2019

Abstract

The following paper presents the use of zero-truncated Poisson model in road accidents analysis. To estimate coefficients, the maximum likelihood method was involved by maximizing the logarithm of the likelihood function, which is formulated as a joint density function for all observations. Also, the *general-to-specific* model selection was proposed to find a few relevant variables from a large list of potentially important characteristics.

The main hypothesis to prove was whether (and the second one- to what extent) any of proposed road and weather conditions has an impact on the number of vehicles involved in accidents. The basis of following investigation was database shared by Department for Transport in The United Kingdom. The database consists of detailed road safety data about the circumstances of personal injury road accidents in 2017. Conclusions from the following research can be helpful for the government to increase public roads safety.

1. Wstęp

Na drogach co roku giną lub doznają poważnych obrażeń miliony osób na całym świecie. Organizacja WHO szacuje, że wypadki drogowe są piątą co do wielkości, globalną przyczyną zgonów. Właśnie dlatego niezwykle ważnym aspektem poprawy bezpieczeństwa ruchu drogowego są liczne badania wypadków drogowych pod kątem ich przyczyn. W niniejszym badaniu obserwacji poddano wpływ różnych okoliczności wypadków na liczbę aut, które w nich uczestniczą.

Badania wypadków drogowych zazwyczaj bazują na danych zagregowanych na poziomie dziennym, miesięcznym lub rocznym. Dowiedziono jednak, że modelowanie zagregowanych danych dotyczących wypadków może prowadzić do powstania błędu ekologicznego, polegającego na nieuzasadnionym przenoszeniu wniosków z estymacji grupowych na zależności na poziomie jednostkowym[1]. W celu wyeliminowania powyższego zagrożenia w niniejszym badaniu wykorzystano zdezagregowane dane dotyczące każdego pojedynczego wypadku[2] mającego miejsce w Wielkiej Brytanii w 2017 roku.

Ze względu na nieujemne częstości wypadków drogowych, model licznikowy z rozkładem Poissona jest jednym z częściej wykorzystywanych modeli w analizach wypadków[3].

[1] Qin, X., Ivan, J.N., Ravishanker, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accid. Anal. Prev. 36 (2), 183–191.

[2] W Wielkiej Brytanii wypadek definiowany jest jako zdarzenie mające miejsce na drodze publicznej, w którym przynajmniej jeden pojazd został uszkodzony oraz przynajmniej jedna osoba doznała obrażeń lub poniosła śmierć na miejscu.

[3] Xin Pei, N.N. Sze, S.C. Wong, Danya Yao, Bootstrap resampling approach to disaggregate analysis of road crashes in Hong Kong, Accident Analysis and Prevention (p. 512-520) , October 2016

W przypadku liczby aut biorących udział w zdarzeniu mamy jednak do czynienia z wartościami większymi od zera. W związku z tym, na potrzeby niniejszego badania, wykorzystano model ograniczonego Poissona (w wartości 0). Oszacowania wyżej wymienionego modelu posłużą do zbadania wpływu warunków pogodowych, stanu infrastruktury drogowej oraz ograniczonej widoczności na liczbę pojazdów w wypadku, a tym samym na skalę poniesionych szkód. Wnioski mogą posłużyć do podjęcia kroków w celu zapewnienia większego bezpieczeństwa przez odpowiednie organy ustawodawcze a także służby porządkowe odpowiedzialne za utrzymanie dróg krajowych.

2. Dane

Dane, będące podstawą przeprowadzanego badania, pochodzą z bazy udostępnionej przez Brytyjski Departament Transportu[4]. Baza składa się z 129.971 wypadków na drogach publicznych Wielkiej Brytanii odnotowanych przez policję w 2017 roku. W celu odpowiedzenia na pytanie co wpływa na liczbę aut biorących udział w wypadku wykorzystano informacje dotyczące warunków pogodowych, warunków nawierzchni drogowej, widoczności czy infrastruktury drogowej. Dokładny opis zmiennych wykorzystanych w badaniu znajduje się w Tabeli 1. Oznaczenie (B) przy nazwie zmiennej wskazuje na fakt, że jest ona wartością binarną.

Tabela 1 Zmienne użyte w badaniu

Statistic	Opis
Number_of_Vehicles	Liczba pojazdów w wypadku
rt_round (B)	Rondo
rt_slipr (B)	Pas zjazdowy
lc_darkness (B)	Warunki świetlne: ciemno, ograniczona widoczność
wc_rain (B)	Deszcz
sc_signal (B)	Nadzwyczajne warunki drogowe: Niedziałająca sygnalizacja
sc_surfacedef (B)	Nadzwyczajne warunki drogowe: Zły stan nawierzchni
rural (B)	Droga wiejska

3. Metoda

Liczba aut biorących udział w wypadku jest, w niniejszej analizie, wartością, która nie osiąga wartości 0. W związku z tym przyjęto, iż omawiane dane charakteryzują się rozkładem obciętego Poissona:

$$\mathrm{Prob[Y = y|y > 0]} = \frac{(e^{-\lambda}\lambda^{y})/y!}{\mathrm{Prob[Y>0]}} = \frac{(e^{-\lambda}\lambda^{y})/y!}{1-\mathrm{Prob[Y=0]}} = \frac{(e^{-\lambda}\lambda^{y})/y!}{1-e^{-\lambda}}, \lambda > 0, y \qquad (1)$$

W celu oszacowania modelu wykorzystano metodę największej wiarygodności (MNW). Metoda ta pozwala na oszacowanie parametrów poprzez maksymalizację funkcji gęstości dla całej próby. W celu uproszczenia obliczeń, zamiast funkcji gęstości stosuje się jej monotoniczne przekształcenie, czyli logarytmowanie. Zlogarytmowana funkcja wiarygodności w omawianym modelu przybiera zatem następującą postać :

$$ln\, L = -\sum \lambda + \sum y \ln \lambda - \sum \ln 1 - e^{-\lambda}\,, \qquad (2)$$

[4] data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

gdzie parametr $\lambda = e^{\beta X}$ i β jest wektorem oszacowań, natomiast X potencjalnych czynników wpływających na liczbę aut.

Do maksymalizacji funkcji log-wiarygodności (2) zastosowano metodę Newtona Rapsona, dostępną w pakiecie statystycznym R wraz z biblioteką *maxLik.* W tym celu użyto analityczną postać gradientu (3) oraz hesjanu (4) funkcji gęstości:

$$G_n = \left[\frac{\partial \ln L}{\partial \theta_1} \quad \frac{\partial \ln L}{\partial \theta_2} \quad \dots \quad \frac{\partial \ln L}{\partial \theta_K}\right] \tag{3}$$

$$H_n = \begin{bmatrix} \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_K} \\ \frac{\partial^2 \ln L}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ln L}{\partial \theta_2 \partial \theta_2} & \cdots & \frac{\partial^2 \ln L}{\partial \theta_2 \partial \theta_K} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 \ln L}{\partial \theta_K \partial \theta_1} & \frac{\partial^2 \ln L}{\partial \theta_K \partial \theta_2} & \cdots & \frac{\partial^2 \ln L}{\partial \theta_K \partial \theta_K} \end{bmatrix} \tag{4}.$$

Doboru istotnych zmiennych do modelu dokonano za pomocą procedury od ogółu do szczegółu na podstawie testu LR. Polega ona na iteracyjnym szacowaniu modelu z ograniczeniem $H_0: \beta_i = 0$, gdzie β_i jest parametrem, który na podstawie wartości p-value z poprzedniej iteracji modelu uznany jest za nieistotny.

4. Wyniki

Dobór istotnych zmiennych do modelu został przeprowadzony w dwóch krokach :

1. Oszacowano wstępny model na podstawie którego stwierdzono nieistotność zmiennej rt_slipr:
   $Number\_of\_Vehicles = \beta_0 + \beta_1 rt\_round + \beta_2 rt\_slipr + \beta_3 rural + \beta_4 wc\_rain + \beta_5 lc\_darkness + \beta_6 sc\_signal + \beta_7 sc\_surfacedef.$
2. W celu usunięcia zmiennej z modelu, oszacowano kolejny model z ograniczeniem $H_0: \beta_2 = 0$:
   $Number\_of\_Vehicles = \beta_0 + \beta_1 rt\_round + \beta_2 rt\_slipr + \beta_3 rural + \beta_4 wc\_rain + \beta_5 lc\_darkness + \beta_6 sc\_signal + \beta_7 sc\_surfacedef$.
   Następnie przeprowadzono test LR na poprawność nałożonego ograniczenia. Statystyka wyniosła 0.23, co przy kwantylu rzędu 0.05 rozkładu chi-kwadrat z 1 stopniem swobody równym 3.84 nie dało podstaw do odrzucenia przyjętego ograniczenia. Z modelu 1. usunięto zatem zmienną rt_slipr.

Wyniki oszacowań modeli widoczne są w Tabeli 2. Znajdują się w niej także wyniki testu LR oraz wartości kryterium informacyjnego AIC, które dowodzą słuszności uznania modelu 2 jako lepiej dopasowanego (w przypadku drugiego oszacowania AIC charakteryzuje się mniejszą wartością).

Tabela 2 Wyniki oszacowań modelu 1 i 2

	Tr.Poisson 1	Tr.Poisson 2
int.	0.307860 **	0.307984 **
	(0.004043)	(0.004033)
rt_round	0.061528 **	0.161359 **
	(0.011006)	(0.011015)
rt_slipr	0.012526	
	(0.026019)	
rural	0.094035 **	0.094119 **
	(0.005866)	(0.005852)
wc_rain	-0.040771 **	-0.040772 **
	(0.008950)	(0.008939)
lc_darkness	-0.082877 *	-0.082864 *
	(0.006400)	(0.006384)
sc_signal	0.127010 **	0.127011 **
	(0.048381)	(0.048346)
sc_surfacedef	0.660244 **	0.661936 **
	(0.124042)	(0.102913)
AIC	128944	49429
Log Likehood	-64465	-64465
Num. obs.	129971	129971
LR chi2(1)		0.228493
chi2(1)		3.841459

** p < 0.01, * p < 0.05

5. Wnioski

Na podstawie oszacowań parametrów z Tabeli 1 można zauważyć, iż największy wpływ na liczbę aut biorących udział w wypadku mają zmienne dotyczące nadzwyczajnych warunków infrastruktury drogowej. Awaria sygnalizacji świetlnej w czasie jazdy powoduje, że w wypadku uczestniczy 13% więcej aut. Natomiast poważne uszkodzenie nawierzchni powoduje wzrost pojazdów w wypadku aż o 66%. Powyższe wnioski mogą być niezwykle istotne z punktu widzenia planowania budżetu na służby porządkowe dróg publicznych. Większa niezawodność sygnalizacji świetlnej oraz szybka i efektywna naprawa nawierzchni drogowej mogą znacznie zredukować skalę poniesionych szkód i strat podczas wypadków drogowych.

Utrudniona widoczność oraz deszcz (śliska nawierzchnia) z kolei przekładają się na mniejszą ilość aut biorących udział w wypadku. Wynikać to może z faktu, iż w takich warunkach większość kierowców zwiększa czujność a wypadki są obwarowane mniejszymi stratami.

Na drogach wiejskich zderzeniu ulega 10% więcej pojazdów niż na drogach miejskich. Ronda z kolei są newralgicznymi elementami dróg, na których zderza się ze sobą 16% więcej pojazdów. Z pewnością warto wziąć powyższe dwa oszacowania pod uwagę podczas planowania budowy nowych odcinków drogowych oraz planowania ograniczeń prędkości poza obszarem miejskim.

6. Aneks

```
###import danych###
ACC= read.csv(file ="ukwypadki.csv", sep=";",header=TRUE)
### wykorzystane biblioteki ###
library(maxLik)
library(texreg)
############1st model#############
#### Zmienne objasniajace i zm.objasniana ####
z1=ACC$rt_round
z2=ACC$rt_slipr
z3=ACC$rural
z4=ACC$wc_rain
z5=ACC$lc_darkness
z6=ACC$sc_signal
z7=ACC$sc_surfacedef
vv=data.frame(z1,z2,z3,z4,z5,z6,z7)
y=ACC$Number_of_Vehicles

#(-1-)###Log-Likehood function####
LL=function(parametry) {
  #beta
  B0 = parametry[1]
  B1 = parametry[2]
  B2 = parametry[3]
  B3 = parametry[4]
  B4 = parametry[5]
  B5 = parametry[6]
  B6 = parametry[7]
  B7 = parametry[8]
  #vars
  z1=ACC$rt_round
  z2=ACC$rt_slipr
  z3=ACC$rural
  z4=ACC$wc_rain
  z5=ACC$lc_darkness
  z6=ACC$sc_signal
  z7=ACC$sc_surfacedef
  #lambda:
  xb = B0+B1*z1+B2*z2+B3*z3+B4*z4+B5*z5+B6*z6+B7*z7
  lambda = exp(xb)
  #log-likelihood
  return(-sum(lambda)+sum(y*log(lambda))-sum(log(1-exp(-lambda))))
}

#(-2-)###GRADIENT function####
gradient=function(parametry){
  #beta
  B0 = parametry[1]
  B1 = parametry[2]
  B2 = parametry[3]
  B3 = parametry[4]
  B4 = parametry[5]
  B5 = parametry[6]
  B6 = parametry[7]
  B7 = parametry[8]
  #lambda
  lambd=z1*B1+B0+B2*z2+B5*z5+B4*z4+z3*B3+B7*z7+B6*z6
  #declare gr as vector
  gr=rep(0,8)
  #gradient function
  gr[1]=sum(y/(lambd))-sum(exp(-lambd)/(1-exp(-lambd)))-1
  for (i in 2:8){
    gr[i]=sum(vv[,i-1]*y/(lambd))-sum((vv[,i-1]*exp(-lambd))/(1-exp(-lambd)))-sum(vv[,i-1])
  }
  gr
  return(gr)
}
```

```
#(-3-)###HESSJAN function####
hessjan= function(parametry){
 #beta
 B0 = parametry[1]
 B1 = parametry[2]
 B2 = parametry[3]
 B3 = parametry[4]
 B4 = parametry[5]
 B5 = parametry[6]
 B6 = parametry[7]
 B7 = parametry[8]
 #lambda
 lambd=z1*B1+B0+B2*z2+B5*z5+B4*z4+z3*B3+B7*z7+B6*z6
 #declare h as matrix
 h=matrix(nrow=8, ncol=8)
 #hessjan function
 #1.col from 2. to 8. row
 for (k in 2:8){
  h[k,1]=sum(vv[,k-1]*exp(-lambd)/(1-exp(-lambd)))+sum(vv[,k-1]*exp(-2*(lambd))/
                        (1-exp(-lambd))^2)-sum(vv[,k-1]*y/(lambd)^2)
 }
 h
 ##1.row##
 #d2LL/dB0dB0
 h[1,1]=sum(exp(-lambd)/(1-exp(-lambd)))+sum(exp(-2*(lambd))/
                (1-exp(-lambd))^2)-sum(y/(lambd)^2)
 #d2LL/dB0dB1
 h[1,2]=-sum(z1*y/(lambd)^2)+sum(z1*exp(-lambd)/(1-exp(-lambd)))+
  sum(z1*exp(-2*(lambd))/(1-exp(-lambd))^2)
 #3. to 8. col
 for (j in 3:8){
  h[1,j]=sum(vv[,j-1]*exp(-lambd)/(1-exp(-lambd)))+sum(vv[,j-1]*exp(-2*(lambd))/
                        (1-exp(-lambd))^2)-sum(vv[,j-1]*y/(lambd)^2)
 }
 h
 ##2. to 8. row##
 for(q in 2:8){
  for (l in 2:8){
   h[q,l]=sum(vv[,q-1]*vv[,l-1]*exp(-lambd)/(1-exp(-lambd)))+sum(vv[,q-1]*vv[,l-1]*exp(-2*(lambd))/
       (1-exp(-lambd))^2)-sum(vv[,q-1]*vv[,l-1]*y/(lambd)^2)
  }
 }
 h
 return(h)
}

####1st model estimation####
max_beta <- maxLik(logLik=LL,grad=gradient,hess=hessjan,
start=c(int.=0.1,rt_round=0.1,rt_slipr=0.1,rural=0.1,wc_rain=0.1,lc_darkness=0.1,sc_signal=0.1,sc_surfacedef=0.1))
summary(max_beta)
m1=summary(max_beta)
aic1=AIC(max_beta)
```

```
############2nd model############
#### Zmienne objasniajace i zm.objasniana ####
z1=ACC$rt_round
z3=ACC$rural
z4=ACC$wc_rain
z5=ACC$lc_darkness
z6=ACC$sc_signal
z7=ACC$sc_surfacedef
vv=data.frame(z1,z3,z4,z5,z6,z7)
y=ACC$Number_of_Vehicles
#(-1.1-)###Log-Likehood function####
LL2=function(parametry) {
 #bety
 B0 = parametry[1]
 B1 = parametry[2]
 B3 = parametry[3]
 B4 = parametry[4]
 B5 = parametry[5]
 B6 = parametry[6]
 B7 = parametry[7]
 #lambda:
 xb = B0+B1*z1+B3*z3+B4*z4+B5*z5+B6*z6+B7*z7
 lambda = exp(xb)
 #logarytm f.wiarygodnosci
 return(-sum(lambda)+sum(y*log(lambda))-sum(log(1-exp(-lambda))))
}

#(-2.1-)###GRADIENT function####
gradient2=function(parametry){
 #beta
 B0 = parametry[1]
 B1 = parametry[2]
 B3 = parametry[3]
 B4 = parametry[4]
 B5 = parametry[5]
 B6 = parametry[6]
 B7 = parametry[7]
 #lambda
 lambd=z1*B1+B0+B5*z5+B4*z4+z3*B3+B7*z7+B6*z6
 #declare gr as vector
 gr=rep(0,7)
 #gradient function
 gr[1]=sum(y/(lambd))-sum(exp(-lambd)/(1-exp(-lambd)))-1
 for (i in 2:7){
  gr[i]=sum(vv[,i-1]*y/(lambd))-sum((vv[,i-1]*exp(-lambd))/(1-exp(-lambd)))-sum(vv[,i-1])
 }
 gr
 return(gr)
}
#(-3.1-)###HESSJAN function####
hessjan2= function(parametry){
 B0 = parametry[1]
 B1 = parametry[2]
 B3 = parametry[3]
 B4 = parametry[4]
 B5 = parametry[5]
 B6 = parametry[6]
 B7 = parametry[7]
 #lambda
 lambd=z1*B1+B0+B5*z5+B4*z4+z3*B3+B7*z7+B6*z6
 #zadeklarowanie hessj. jako matrix
 h=matrix(nrow=7, ncol=7)
 #funkcja hessjan
 #1.col from 2nd row
 for (k in 2:7){
  h[k,1]=sum(vv[,k-1]*exp(-lambd)/(1-exp(-lambd)))+sum(vv[,k-1]*exp(-2*(lambd))/
                                        (1-exp(-lambd))^2)-sum(vv[,k-1]*y/(lambd)^2)
 }
 h
 ##1.row##
 #d2LL/dB0dB0
```

```
  h[1,1]=sum(exp(-lambd)/(1-exp(-lambd)))+sum(exp(-2*(lambd))/
                                  (1-exp(-lambd))^2)-sum(y/(lambd)^2)
  #d2LL/dB0dB1
  h[1,2]=-sum(z1*y/(lambd)^2)+sum(z1*exp(-lambd)/(1-exp(-lambd)))+
    sum(z1*exp(-2*(lambd))/(1-exp(-lambd))^2)
  #3 to 7 col
  for (j in 3:7){
    h[1,j]=sum(vv[,j-1]*exp(-lambd)/(1-exp(-lambd)))+sum(vv[,j-1]*exp(-2*(lambd))/
                                        (1-exp(-lambd))^2)-sum(vv[,j-1]*y/(lambd)^2)
  }
  h
  ##2. to 7. row##

  for(q in 2:7){
    for (l in 2:7){
      h[q,l]=sum(vv[,q-1]*vv[,l-1]*exp(-lambd)/(1-exp(-lambd)))+sum(vv[,q-1]*vv[,l-1]*exp(-2*(lambd))/
                                                (1-exp(-lambd))^2)-sum(vv[,q-1]*vv[,l-1]*y/(lambd)^2)
    }
  }
  h
  return(h)

}

####2nd model estimation####
max_beta2 <- maxLik(logLik=LL2,grad=gradient2,hess=hessjan2,
                    start=c(int.=0.1,rt_round=0.1,rural=0.1,wc_rain=0.1, lc_darkness=0.1,sc_signal=0.1,sc_surfacedef=0.1))
summary(max_beta2)
m2=summary(max_beta2)
aic2=AIC(max_beta2)

####################
#PUBLICATION TABLE#
####################
#m1 table
m1_est=(as.table(m1$estimate))
#m2 table
m2_est=(as.table(m2$estimate))
###
#aic
aic1=AIC(max_beta1)
aic2=AIC(max_beta2)
#log-likelihood
llik1=m1$loglik
llik2=m2$loglik
#num.obs
n=length(ACC[,1])
#LR test with h0:B2=0
LRstat= 2*(llik1-llik2)
print(LRstat)
#kwantyl rozkladu chi-kwadrat z 1d.f. dla p=0.95
chi2_1=qchisq(p=0.95,df=1)
#brak podst do odrzucenia h0:B0=0

#create texreg objects
library(texreg)
m2_tr=createTexreg(coef.names =row.names(m2_est),coef=m2_est[,1],m2_est[,2],pvalues =m2_est[,4],
              gof.names = c("AIC","Log Likehood","Num. obs.","LR chi2(1)", "chi2(1)" ),
              gof=c(aic1,llik1,n,LRstat,chi2_1) ,gof.decimal=c(F,F,F,T,T))
m1_tr=createTexreg(coef.names =row.names(m1_est),coef=m1_est[,1],m1_est[,2],pvalues =m1_est[,4],
              gof.names = c("AIC","Log Likehood","Num. obs."),gof=c(aic2,llik2,n),
              gof.decimal=c(F,F,F) )

#publication table with 2 models#
screenreg(list(m1_tr,m2_tr),file=NULL,stars=c(0.01,0.05),custom.model.names = c("Tr.Poisson 1", "Tr.Poisson 2"),digits=6)
```