Regresja logistyczna na przykładzie katastrofy Challengera

Elżbieta Kukla

24 lutego 2011

Spis treści

Źródła:

- *http://data.princeton.edu/wws509/notes/a2.pdf*
- *http://www.public.iastate.edu/ stat415/stephenson/stat415_chapter3.pdf*
- *http://www.stat.ucdavis.edu/ johnson/st145/Ch5.pdf*
- *http://www.mimuw.edu.pl/ pokar/Rabczenko/FarawayExtendLinMod06.pdf*

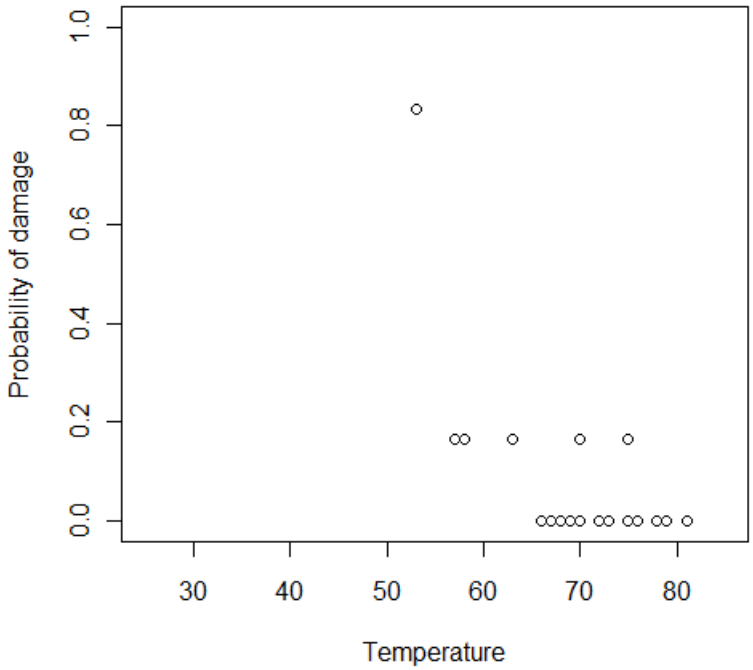
1 Katastrofa Challenger'a

28 stycznia 1986 roku miała miejsce katastrofa amerykańskiego promu kosmicznego Challenger. Katastrofa ta miała miejsce 73 sekundy po starcie wahadłowca. W celu zbadania przczyn katastrofy przeprowadzono szczegółowe badania, w czasie których zwrócono uwagę na gumowe pierścienie (o okrągłym przekroju) uszczelniające dodatkowe rakiety. Rakiety te składają się z ładunku paliwa i silnika rakietowego na paliwo stałe. Wahadłowiec składa się z dwóch takich rakiet, każda posiada trzy takie pierścienie. Przez pierwsze dwie minuty lotu działają one równolegle z głównymi silnikami promu, pozwalając na pokonanie grawitacji Ziemi oraz przyspieszanie. Na wysokości około 45 kilometrów oba dodtkowe silniki odłączają się od zewnętrznego zbiornika, opadają na spadochronach i wodują w Oceanie Atlantyckim. Są wyławiane przez statki i transportowane na ląd, gdzie są przystosowywane do ponownego użycia.

Stwierdzono, że w niskiej temperaturze guma staje się bardziej łamliwa i jest mniej skutecznym szczeliwem, a przyczyną katastrofy było uszkodzenie pierścienia uszczelniającego w prawym silniku wspomagającym, które nastąpiło najprawdopodobniej między pierwszą a trzecią sekundą lotu. Na skutek tego uszkodzenia i w efekcie oddziaływania gorących gazów wewnątrz silnika na powstałą nieszczelność, na zewnątrz połączenia pojawił się płomień. Płomień ten przepalił dziurę w zbiorniku zewnętrznym wahadłowca, co spowodowało eksplozję tego zbiornika i zniszczenie całego promu.

W momencie startu temperatura powietrza wynosiła 31^oF, czyli trochę poniżej 0^oC. Zadawano sobie pytanie, czy można było przewidzieć katastrofę? Zaczęto badać 23 poprzednie misje wahadłowców, dla których istniały dane (łącznie odbyły 24 loty, lecz raz rakiet wspomagających nie odnaleziono), zostały odnotowane pewne oznaki zniszczenia na niektórych pierścieniach. Dla każdej z tych misji, znana jest temperatura powietrza, przy której startował prom oraz liczba pierścieni (spośród sześciu) wykazujących pewne uszkodzenia.

2 Model liniowy

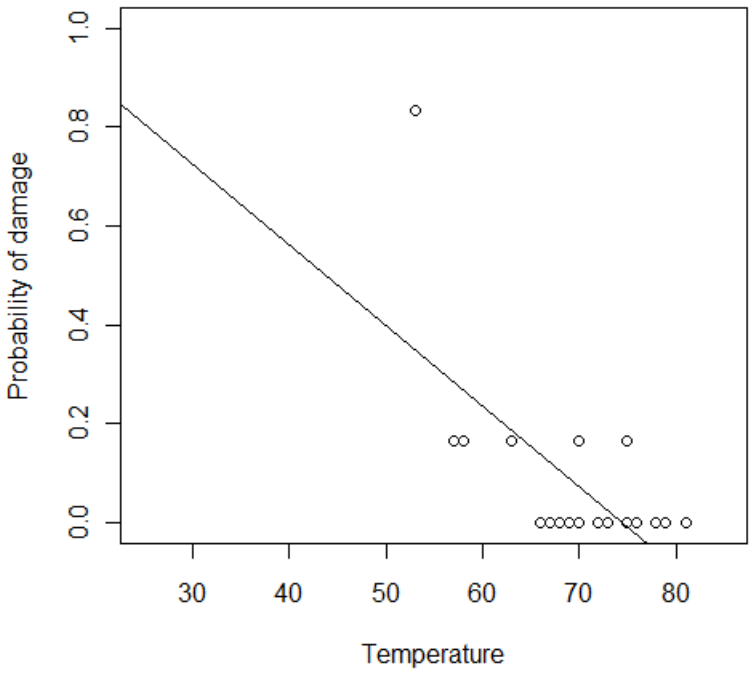


Rysunek 1: R - proporcja zniszczonych pierścieni w zależności od temperatury.

Interesuje nas teraz, jak prawdopodobieństwo uszkodzenia danego pierścienia jest związane z tempreturą powietrza w chwili startu i przewidzenie tego prawdopodobieństwa, gdy temperatura wynosi 31^oF. Najprostsze podejście oparte na linowym modelu, po prostu dopasowuje prostą do tych danych.

Przy tym podejściu napotykamy wiele problemów, co widać na wykresie. Przewidywane wartości prawdopodobieństwa mogą być większe od jedynki lub mniejsze od zera. Ktoś może zasugerować obcięcie tych wartości do przedziału $[0, 1]$, ale nie wydaje się, że to jest dobry sposób.

Lepiej założyć, że liczba zniszczeń ma rozkład dwumianowy (Bernoulliego). W przypadku modelu liniowego, wymagamy aby błędy miały rozkład normalny (do dokładnego testowania). Jednak, w przypadku rozkładu dwumianowego z jedynie 6 próbami, przybliżenie rozkładu normalnego byłoby naciągane. W dodatku wariancja zmiennej o rozkładzie Bernoulliego nie jest stała (jest funkcją prawdopodobieństwa p_i), co nie spełnia kolejnego istotnego założenia modelu liniowego. Ewidentnie standardowy liniowy model nie jest tutaj odpowiedni. Chociaż, możemy próbować naprawić niektóre z tych problemów (transformacje itp.), lepiej wprowadzić inny model, który dokładnie odpowiada danym o rozkładzie dwumianowym.



Rysunek 2: Zniszczone pierścienie w 23 misjach wahadłowca jako funkcja temperatury startu. Prosta otrzymana przy użyciu MNK.

3 Wstęp do regresji logistycznej

Przypuśćmy, że zmienna odpowiedzi Y_i dla $i = 1, \ldots, n_i$ ma rozkład Bernoulliego z parametrami $B(n_i, p_i)$, tak że

$$P(Y_i = y_i) = \binom{n}{k} p_i^{y_i} (1 - p_i)^{n_i - y_i}.$$

Dalej, załóżmy, że zmienne Y_i są niezależne. Pojedyncza próba, z których składa się Y_i, zależy od tych samych q predyktorów $(x_{i1}, \ldots, x_{iq})$. Grupa prób nazywana jest *covariate class*. Potrzebujemy modelu, który opisuje relacje $x_1, \ldots, x_q$ w stosunku do p. Postępując tak jak w modelu liniowym, konstruujemy liniowy predyktor:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}.$$

Jako że liniowy predyktor może mieścić zarówno jakościowe jak i ilościowe predyktory przy użyciu sztucznych zmiennych (ang. *dummy*) oraz pozwala na transformacje i kombinacje oryginalnych predyktorów, jest on bardzo elastyczny. Fakt, że możemy wyrazić efekty predyktorów na zmienną odpowiedzi wyłącznie przez liniowy predyktor jest ważny. Ta idea może być rozszerzona do modeli o innych typach zmiennych odpowiedzi i jest jedną z ważnych cech szerszej klasy uogólnionych modeli liniowych omówionej przy okazji uogólnionych

modeli liniowych (GLM).

Już powyżej zaobserwowaliśmy, że ustalenie $\eta_i = p_i$ nie jest odpowiednie, ponieważ chcemy aby $0 \leqslant p_i \leqslant 1$. Zamiast tego powinniśmy użyć funkcji wiążącej g takiej że $\eta_i = g(p_i)$. Do tego potrzebujemy funkcji g – monotonicznej i takiej że $0 \leqslant g^{-1}(\eta) \leqslant 1$ dla dowolnego η.
Mamy trzy popularne funkcje:

1. Logit: $\eta = \log(p/(1-p))$
2. Probit: $\eta = \phi^{-1}(p)$, gdzie ϕ^{-1} jest odwrotnością dystrybuanty rozkładu normalnego
3. $log - log$: $\eta = \log(-log(1-p))$ (ang. *complementary log-log*). Idea użycia funkcji wiążącej jest jedną z głównych idei uogólnionych modeli liniowych.

4 Rozkład Bernoulliego – wyprowadzenie funkcji logit

Dystrybuanta rozkładu Bernoulliego:

$$P(Y_i = y_i) = f_i(y_i) = \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$$

gdzie
y_i – liczba zniszczonych pierścieni po i-tej misji wahadłowca
n_i – łączna liczba pierścieni w obu rakietach (u nas zawsze 6)
p_i – obliczona proporcja zniszczonych pierścieni Kolejne przekształcenia:

$$\log f_i(y_i) = y_i \log(p_i) + (n_i - y_i)\log(1-p_i) + \log \binom{n_i}{y_i}$$

$$\log f_i(y_i) = y_i \log(\frac{p_i}{1-p_i}) + n_i \log(1-p_i) + \log \binom{n_i}{y_i}$$

Mamy rodzinę wykładniczą, bo powyższe wyrażenie ma postać:

$$\log f_i(y_i) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\theta)} + c(y_i, \theta).$$

Teraz zauważamy, patrząc na współczynnik przy y_i, że kanonicznym parametrem jest logit p_i:

$$\theta_i = \log(\frac{p_i}{1-p_i}) = \eta_i.$$

Rozwiązując to dla p_i, dostajemy $p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$, więc $1 - p_i = \frac{1}{1+e^{\theta_i}}$.
Stąd łatwo zauważyć, że $b(\theta_i) = n_i \log(1 + e^{\theta_i})$ oraz $c(y_i, \phi) = \log \binom{n_i}{k}$.
Przyjmujemy, że $a_i(\phi) = \phi$ i $\phi = 1$.
Teraz łatwo też sprawdzić wartość oczekiwaną i wariancję:

- $E(Y_i) = \mu_i = b'(\theta_i) = n_i \frac{e^{\theta_i}}{1+e^{\theta_i}} = n_i p_i$
- $Var(Y_i) = v_i = a_i(\phi) b''(\theta_i) = n_i \frac{e^{\theta_i}}{(1+e^{\theta_i})^2} = n_i p_i (1 - p_i)$.

5 Estymacja metodą największej wiarogodności

Najpierw oszacujmy parametry modelu, użyjemy do tego metody największej wiarogodności. Logarytm funkcji wiarogodności jest dany przez:

$$l(\beta) = \sum_{i=1}^{n} [y_i \eta_i - n_i \log(1 + e_i^{\eta}) + \log \binom{n_i}{y_i}].$$

Powinniśmy zmaksymalizować to wyrażenie w celu otrzymania estymatorów $\hat{\beta}$ i użyć standardowej teorii do obliczenie przybliżonych standardowych błędów. Zmaksymalizowanie tego wyrażenia nie jest w tym przypadku takie proste. Estymatory największej wiarogodności mogą być łatwo i dokładnie znajdowane analitycznie w przypadku uogólnionych modeli liniowych tylko wtedy, gdy mamy rozkład normalny. Zazwyczaj musimy używać optymalizacji numerycznej. Stosujemy metodę Newtona-Raphsona ze scoringiem Fishera. W 1989 roku McCullagh i Nelder pokazali, że ta optymalizacja jest równoważna iterowanej ważonej metodzie najmniejszych kwadratów (ang. *IRWLS – iteratively reweighted least squares*).

6 Analiza w R

6.1 Logit

W naszym modelu mamy oczywiście tylko jedną zmienną objaśniającą – temperaturę powietrza, dlatego też mamy:

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_i.$$

Użyjemy R, żeby estymować regresyjne parametry dla danych Challengera.
Dla zmiennej odpowiedzi o rozkładzie Bernoulliego, potrzebujemy dwóch informacji o wartościach odopwiedzi – y i n. W R jednym ze sposobów osiągnięcia tego, jest utworzenie dwukolumnowej macierzy z pierwszą kolumną reprezentującą liczbę sukcesów y i drugą kolumną z liczbą porażek $n - y$. Określiliśmy, żę zmienna odpowiedzi ma rozkład Bernoulliego. Naturalnym wyborem funkcji wiążącej jest logit (regresja logistyczna) – inne wybory funkcji muszą być specjalnie określone.
Otrzymane współczynniki regresji wynoszą: $\hat{\beta}_0 = 11.6630$ i $\hat{\beta}_1 = -0.2162$, wraz z ich odpowiednimi błędami (o rozkładzie normalnym).

Pokażemy jak wygląda dopasowanie logitu do danych. Zauważmy, jak dopasowanie logitem zbiega asymptotycznie do 0 dla wysokich temperatur oraz do 1 dla niskich. Dopasowane wartości jednak nigdy nie osiągną zera ani jedynki, więc model nigdy nie przewidzi zdarzenia z całkowitą pewnością.

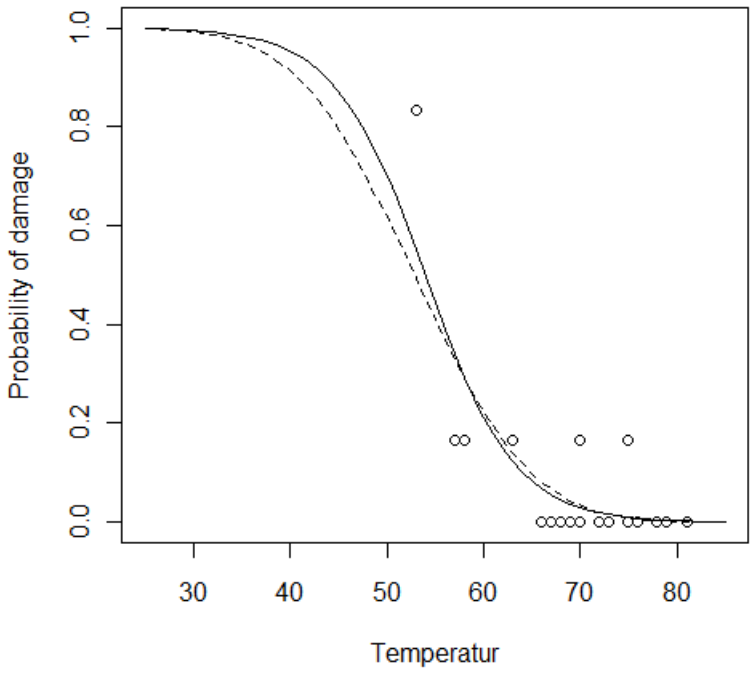
6.2 Probit

Otrzymane współczynniki regresji wynoszą: $\hat{\beta}_0 = 5.5915$ i $\hat{\beta}_1 = -0.1058$, wraz z ich odpowiednimi błędami (o rozkładzie normalnym).

6.3 Porównanie funkcji logit i probit

Chociaż współczynniki wydają się być zupełnie inne, dopasowanie jest podobne, szczególnie w widocznym zakresie temperatur.

Możemy łatwo przewidzieć wartość prawdopodobieństwa w temperaturze 31^oF dla obu modeli: 0.99304 (logit) oraz 0.9896 (probit).



Rysunek 3: Dopasowanie do danych Challengera przy użyciu funkcji logit (linia ciągła) i probit (linia przerywana).

Widzimy bardzo wysokie prawdopodobieństwo zniszczenia w każdym modelu, chociaż musimy rozwinąć techniki testowania, nim wyciągniemy ostateczne wnioski.

A Dokumentacja zbioru danych orings {faraway}

Space Shuttle Challenger O-rings

Description

The 1986 crash of the space shuttle Challenger was linked to failure of O-ring seals in the rocket engines. Data was collected on the 23 previous shuttle missions. The launch temperature on the day of the crash was 31F.

Usage

data(orings)

Format

A data frame with 23 observations on the following 2 variables.

temp – temperature at launch in degrees F

damage – number of damage incidents out of 6 possible

Source

Presidential Commission on the Space Shuttle Challenger Accident, Vol. 1, 1986: 129-131.

References

S. Dalal, E. Fowlkes and B. Hoadley (1989) "Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure." Journal of the American Statistical Association. 84: 945-957.

	temp	damage
1	53	5
2	57	1
3	58	1
4	63	1
5	66	0
6	67	0
7	67	0
8	67	0
9	68	0
10	69	0
11	70	1
12	70	0
13	70	1
14	70	0
15	72	0
16	73	0
17	75	0
18	75	1
19	76	0
20	76	0
21	78	0
22	79	0
23	81	0

B Skrypty w R

```
#proporcja zniszczonych pierścieni w zależności od temperatury
library(faraway)
data(orings)
orings
plot(damage/6~temp,orings,xlim=c(25,85), ylim=c(0,1,
xlab="Temperatura", ylab="P-stwo zniszczenia")

#najprostsze podejście – model linowy
lmod<-lm(damage/6~temp,orings)
abline(lmod)
summary(lmod)

#logit
logitmod<-glm(cbind(damage,6-damage)~temp, family=binomial, orings)
summary(logitmod)
 plot(damage/6~temp, orings, xlim=c(25,85), ylim=c(0,1),
xlab="Temperatura",ylab="P-stwo zniszczenia")
x<-seq(25,85,1)
lines(x,ilogit(11.6630-0.2162*x))

#probit
probitmod<-glm(cbind(damage,6-damage)~temp,
family=binomial(link=probit), orings)
summary(probitmod)

#logit i probit są podobne
lines(x,pnorm(5.5915-0.1058*x),lty=2)

#prawdopodobieństwo zniszczenia przy temperaturze 31F
ilogit(11.6630-0.2162*31)
pnorm(5.5915-0.1058*31)
```