Matematičke metode u marketingu. Generalizirani linearni model

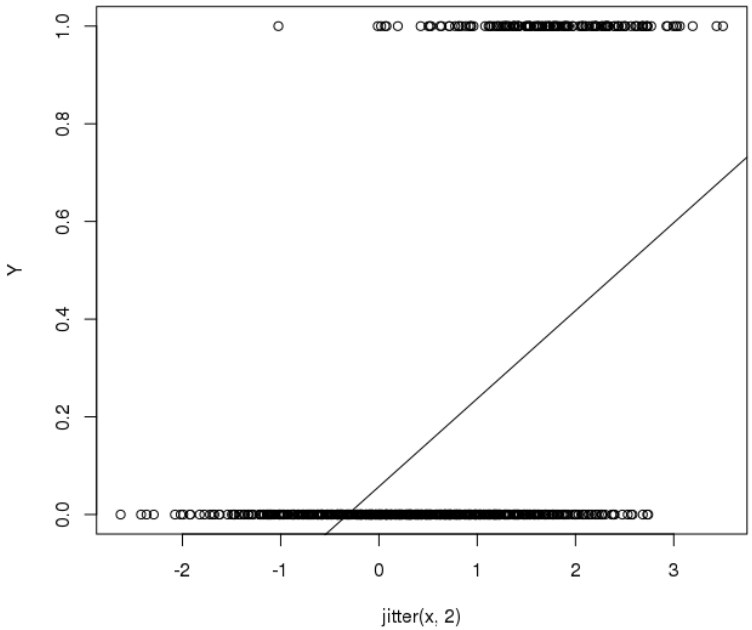
Lavoslav Čaklović
PMF-MO

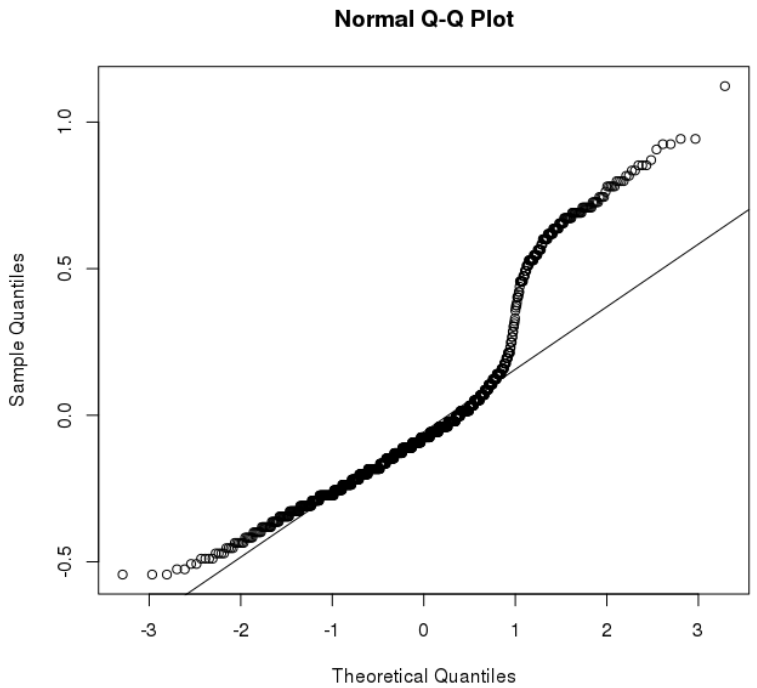
2016

Jedan loš linearni model

$$n = 1000, \quad i = 1, \ldots, n$$

$$Y = \begin{cases} 1 & \text{ako } y_i > 0 \\ 0 & \text{inače} \end{cases} \qquad \begin{aligned} y_i &= -2x_i + rnorm(n) \\ x_i &= round(0.001 * i + rnorm(n), 1) \end{aligned}$$



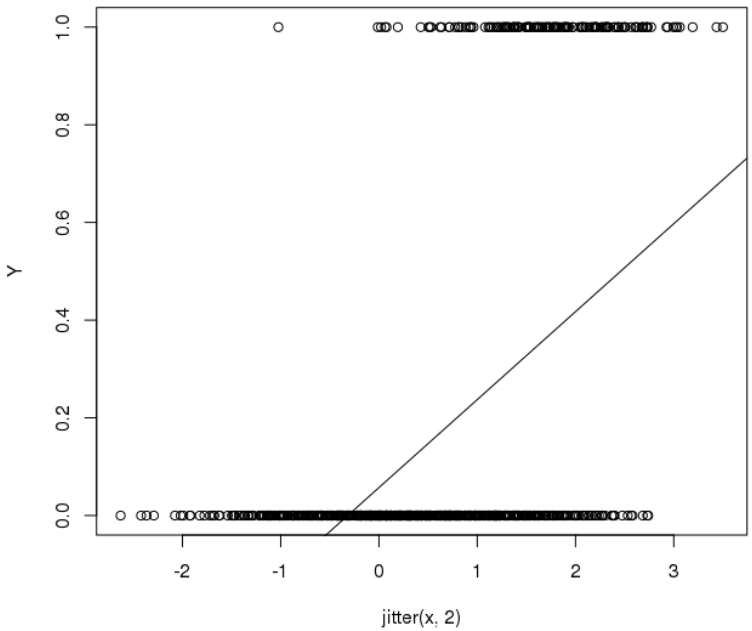


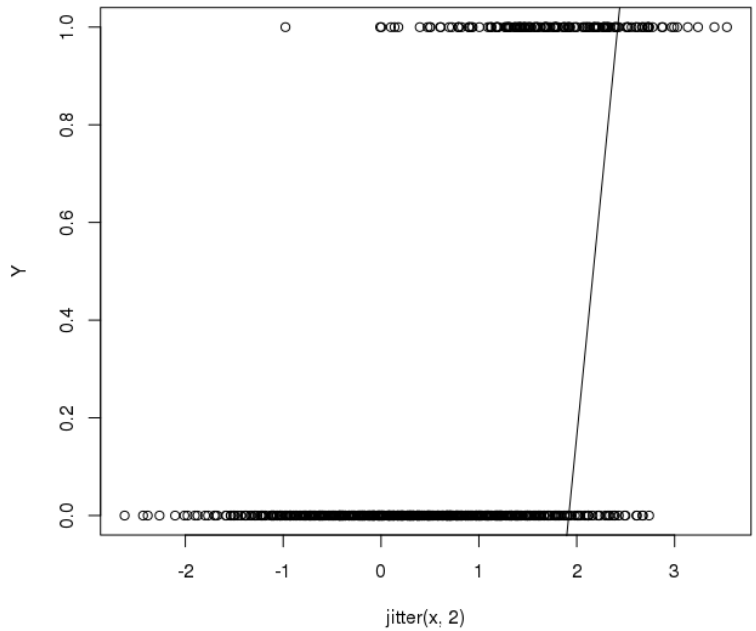
Lijevo: OLS
Desno: qqplot

Jedan loš linearni model

$$n = 1000, \quad i = 1, \ldots, n$$

$$Y = \begin{cases} 1 & \text{ako } y_i > 0 \\ 0 & \text{inače} \end{cases} \qquad \begin{aligned} y_i &= -2x_i + rnorm(n) \\ x_i &= round(0.001 * i + rnorm(n), 1) \end{aligned}$$





Lijevo: OLS
Desno: GLM

Što je uzrok slabe predikcije u modelu ($R^2 = 0.2$)?

Problem je što su varijable $(Y|X = x_i) \sim B(1, p_i)$ gdje p_i ovisi o i pa i njihova varijanca $m_i p_i (1 - p_i)$[1] nije konstantna. To može biti pogubno za linearnu regresiju (vidi `qqplot` na prethodnoj slici desno).

Nešto bolji model

Jednostavnosti radi, promatrajmo jednu dihotomnu varijablu Y i jedan prediktor X. Promatramo

$$p(x) := P(Y = 1|X = x) \tag{1}$$

i regresijski model ($z \in (-\infty, +\infty)$)

$$z := \log \frac{p(x)}{1 - p(x)} = \beta_0 + \beta_1 x + \epsilon. \tag{2}$$

Zamjerka modelu: z ne poprima realne vrijednosti za $p(x) = 0$ i $p(x) = 1$. Ako je $z \in \mathbb{R}$ i p_i u manjoj mjeri ovisno o i nema razloga ne koristiti model (2).

[1] m_i je učestalost podatka x_i.

Što je uzrok slabe predikcije u modelu ($R^2 = 0.2$)?

Problem je što su varijable $(Y|X = x_i) \sim B(1, p_i)$ gdje p_i ovisi o i pa i njihova varijanca $m_i p_i (1 - p_i)$[1] nije konstantna. To može biti pogubno za linearnu regresiju (vidi `qqplot` na prethodnoj slici desno).

Nešto bolji model

Jednostavnosti radi, promatrajmo jednu dihotomnu varijablu Y i jedan prediktor X. Promatramo

$$p(x) := P(Y = 1|X = x) \tag{1}$$

i regresijski model ($z \in (-\infty, +\infty)$)

$$z := \log \frac{p(x)}{1 - p(x)} = \beta_0 + \beta_1 x + \epsilon. \tag{2}$$

Zamjerka modelu: z ne poprima realne vrijednosti za $p(x) = 0$ i $p(x) = 1$. Ako je $z \in \mathbb{R}$ i p_i u manjoj mjeri ovisno o i nema razloga ne koristiti model (2).

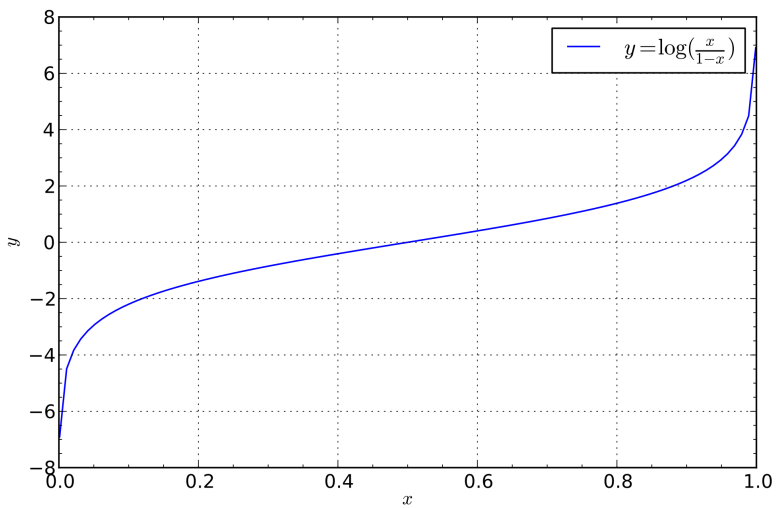
[1] m_i je učestalost podatka x_i.

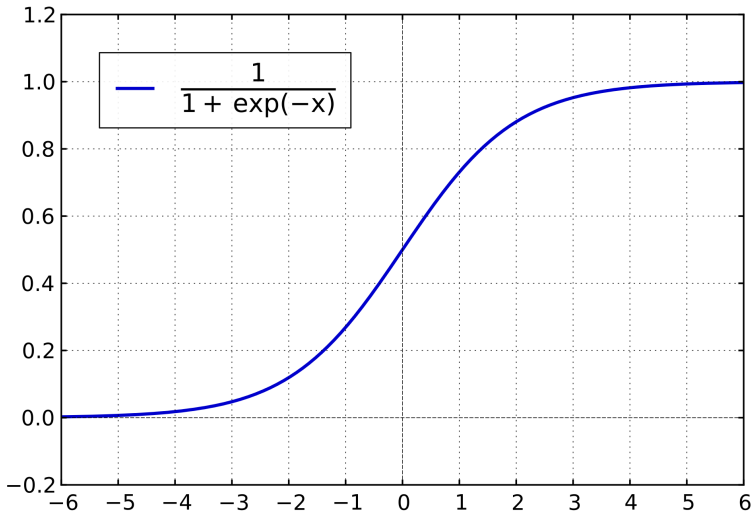
Neki osnovni pojmovi. Fukciju $p \mapsto \log(\frac{p}{1-p})$ nazivamo *logit* funkcijom i u regresiji igra ulogu *link*-funkcije. Njoj inverzna (tzv. *logistička funkcija*) je

$$\Lambda(z) = \frac{e^z}{1+e^z}. \tag{3}$$

Očito:

$$\Lambda'(z) = \Lambda(z)(1-\Lambda(z)). \tag{4}$$





Lijevo: *logit*
Desno: *logistic*

Logistička regresija

Ideja *logističke regresije* je da se umjesto modela (2) i metode najmanjih kvadrata promatra model

$$z := \log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x$$

(bez slučajne greške ϵ) uz pretpostavku da je odzivna varijabla binomna, tj.

$$Y \sim B(1, p).$$

Želimo procijeniti $p(x_i) := P(Y|X = x_i)$ kao funkciju od $\theta = (\beta_0, \beta_1)$.
Koeficijenti modela procjenjuju se maksimizacijom vjerodostojnosti[2]:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} P(Y = y_i|X = x_i) = \prod_{i=1}^{n} p(x_i, \theta)^{y_i}(1 - p(x_i, \theta)^{1-y_i}) \qquad (5)$$

gdje $y_i \in \{0, 1\}$, a

$$p(x, \theta) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} = \Lambda(\beta_0 + \beta_1 x). \qquad (6)$$

[2]Vjerodostojnost danog uzorka je produkt individualnih vjerojatnosti.

Umjesto da maksimiziramo $\mathcal{L}$ maksimizirat ćemo njen logaritam. Zbog jednostavnosti računa uvedimo oznake $p_i = p(x_i, \theta) = \Lambda(z_i)$ i $p' = \frac{d}{dz}\Lambda(z)$:

$$log(\mathcal{L}) = \sum_{y_i=1} log(p_i) + \sum_{y_i=0} log(1-p_i)$$

$$\frac{d}{dz}(log(\mathcal{L}(z)) = \sum_{y_i=1} \frac{p_i'}{p_i} - \sum_{y_i=0} \frac{p_i'}{1-p_i} \quad \text{(pa zbog (4) slijedi)}$$

$$= \sum_{y_i=1} \frac{p_i(1-p_i)}{p_i} - \sum_{y_i=0} \frac{p_i(1-p_i)}{1-p_i} = \sum_{y_i=1}(1-p_i) - \sum_{y_i=0} p_i.$$

Nužni uvjeti maksimalnosti su:

$$\frac{\partial \log(\mathcal{L})}{\partial \beta_0} = \sum_{y_i=1}(1-p_i)\frac{\partial z_i}{\partial \beta_0} - \sum_{y_i=0} p_i \frac{\partial z_i}{\partial \beta_0} = 0$$

$$\frac{\partial \log(\mathcal{L})}{\partial \beta_1} = \sum_{y_i=1}(1-p_i)\frac{\partial z_i}{\partial \beta_1} - \sum_{y_i=0} p_i \frac{\partial z_i}{\partial \beta_1} = 0,$$

odnosno (zbog $z = \beta_0 + \beta_1 x$)

$$\sum_{y_i=1}(1-p_i) - \sum_{y_i=0} p_i = 0$$

$$\sum_{y_i=1}(1-p_i)x_i - \sum_{y_i=0} p_i x_i = 0.$$

Rješenja dobivenih jednadžbi nije moguće dobiti u zatvorenoj formi već se rješavaju numerički Newtonovim iterativnim postupkom. Gornji sustav jednadžbi nema rješenje u slučaju potpune separacije, a za djelomičnu separaciju Newtonova procedura može biti nestabilna.

Za male uzorke, procjena parametara pomoću maksimalne vjerodostojnosti pati od pristranosti.

Primjer (Potpuna separacija)

Najjednostavniji primjer potpune separacije je:

	x	Y
n_1	x_1	1
	$\vdots$	$\vdots$
	x_1	1
n_2	x_2	0
	$\vdots$	$\vdots$
	x_2	0

Nužni uvjet ekstrema vodi na sustav jednadžbi

$$n_1(1-p_1) - n_2 p_2 = 0$$
$$n_1(1-p_1)x_1 - n_2 p_2 x_2 = 0$$

koji ima rješenje samo ako je $x_1 = x_2$.

U slučaju djelomične separacije varijabla Y može u donjem bloku imati nekoliko vrijednosti 1. Potpuna separacija često se javlja kod klasifikacije bolesti koje su učestale sa starošću (prediktor x je vrijeme).

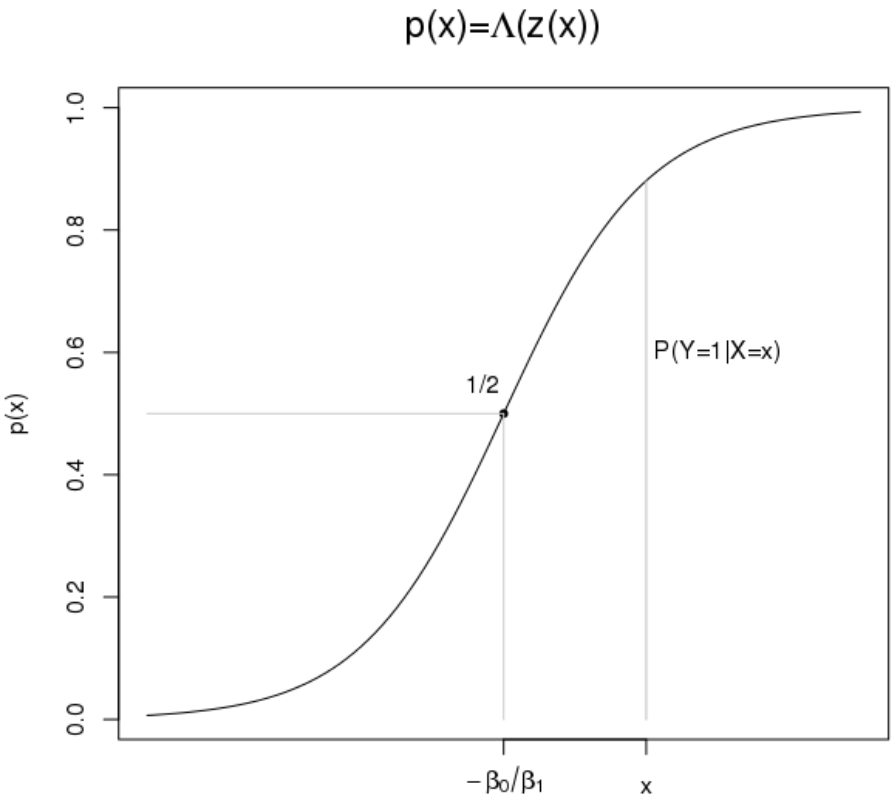
U slučaju separacije svaki softverski paket upozorava korisnika.

Interpretacija modela

Za izračunate β_0, β_1 vjerojatnost $P(X = x)$ i vrijednost od z su:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} = \Lambda(\beta_0 + \beta_1 x)$$

$$z(x) = \log \frac{p(x)}{1 - p(x)} = \beta_0 + \beta_1 x.$$



- Točka infleksije od Λ je $x_i := -\frac{\beta_0}{\beta_1}$, $\Lambda(z_i) = \frac{1}{2}$ (maks. separiranost).
- $\beta_1 = 0 \implies$ omjer ne ovisi o x
- $\exp(\beta_0) \cdot \exp(\beta_1)$ je promjena omjera za jediničnu povećanje x
- Marginalni efekt od p je:
  $p'(x) = \beta_1 \Lambda'(z(x))$
  $= \beta_1 \Lambda(z)(1 - \Lambda(z))$.

Probit model

Funkcija $\Lambda(z)$ je kumulativna distribucija *standardne logističke distribucije* $LOG(0,1)$ čija je gustoća

$$\lambda(z) = \frac{e^z}{(1+z)^2}, \quad \mu = 0, \sigma = \frac{\pi^2}{3}.$$

To sugerira ideju da se umjesto logističke distribucije uzme kumulativna distribucija neke druge slučajne varijable. *Probit* model koristi kumulativnu distribuciju normalne razdiobe:

$$p(x) = \Phi(\beta_0 + \beta_1 x) = \int_{-\infty}^{\beta_0+\beta_1 x} \varphi(x)dx.$$

Marginalni efekt za probit model (tj. derivacija od p) je

$$p'(x) = \frac{\partial \Phi}{\partial x}(x) = \beta_1 \varphi(\beta_0 + \beta_1 x). \tag{7}$$

Logistički model i probit model u praksi nude iste zaključke uz neznatne razlike u procjeni koeficijenata.

Interpretacija marginalnog efekta

Za logit model marginalni efekt

$$p'(x) = \beta_1 \Lambda(z(x))(1 - \Lambda(z(x)))$$

predstavlja promjenu poželjnosti od $Y = 1$ ako se vrijednost prediktora poveća za 1. Primijetimo da je predznak promjene jednak predznaku od β_1.

Za probit model vrijedi isti zaključak (v. formulu (7)).

Kategorijski prediktor

Logistička regresija u R-u.

```
glm(Y ~ x + z, data = data_frame, family = 'binomial')
```

Zadaci

- Prouči R-kôd https://onlinecourses.science.psu.edu/stat504/node/225

Još neke zanimljivosti

- Prvi pokušaj modeliranja binarne ovisnosti: Anscombe, "The transformation of Poisson...", Biometrika (1948).

- U posljednje vrijeme se proučavaju situacije kad je linearni model za binarnu odzivnu varijablu koristan, v. Ottar Hellevik, Linear versus logistic regression when the dependent variable is a dichotomy, Qual Auant (2009)