Pravděpodobnostní metody ve strojírenství

16. Logistická regrese

FAKULTA STROJNÍ ČVUT V PRAZE

prof. RNDr. Gejza Dohnal, CSc.

ak. rok 2023/2024

16. Logistická regrese

Klíčové pojmy:

- Logit
- Šance

Model logistické regrese

$Y \sim \mathrm{Alt}(p)$ $$P(Y = y) = \begin{cases} p & , y = 1, \\ 1 - p & , y = 0, \\ 0 & , \text{jinak} \end{cases}$$ E(Y)=p, Var(Y)=p(1-p)

$$P(Y = y) = p^y(1-p)^{(1-y)}, \quad y \in \{0, 1\}$$

Snažíme se vysvětlit Y pomocí nezávislých náhodných veličin X_1, …, X_k :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \qquad Y \in \{0, 1\}$$

$$P(Y = 1) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \qquad P(Y = 1) \in \langle 0, 1 \rangle$$

Použijeme *šanci* Y (odds): $\mathrm{O}(Y) = \dfrac{P(Y=1)}{P(Y=0)} = \dfrac{P(Y=1)}{1-P(Y=1)} \in (0, \infty)$

a její logaritmus - log-odds: $\ln \dfrac{P(Y=1)}{1-P(Y=1)} \in (-\infty, \infty)$ logit(P($Y = 1$))

$$\ln \frac{P(Y=1)}{1-P(Y=1)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

model logistické regrese

Model logistické regrese

$$\ln \frac{P(Y=1)}{1-P(Y=1)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

odtud:

$$P(Y=1) = \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}}$$

neboli ve vektorovém tvaru:

$$P(Y=1) = \frac{1}{1+\mathrm{e}^{-\mathbf{X}'\vec{\beta}}}$$

kde $\vec{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ a $\mathbf{X} = (1, X_1, \ldots, X_k)'$.

Přesněji:

$$\boxed{P(Y=1|\,\mathbf{X}=\mathbf{x}) = \frac{1}{1+\mathrm{e}^{-\mathbf{x}'\vec{\beta}}}}$$

model logistické regrese

Všimněme si, že je

$$P(Y=0|\,\mathbf{X}=\mathbf{x}) = 1 - P(Y=1|\,\mathbf{X}=\mathbf{x}) = \frac{\mathrm{e}^{-\mathbf{x}'\vec{\beta}}}{1+\mathrm{e}^{-\mathbf{x}'\vec{\beta}}} = \frac{1}{1+\mathrm{e}^{\mathbf{x}'\vec{\beta}}}$$

Model logistické regrese

$$P(Y = 1 | \mathbf{X} = \mathbf{x}) = \frac{1}{1 + \mathrm{e}^{-\mathbf{x}'\vec{\beta}}}$$

Máme-li n pozorování $Y_1, \ldots, Y_n$, veličiny Y a datovou matici $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1, \ldots, \mathbf{X}_n)$, kde každému Y_i odpovídá vektor náhodných veličin $\mathbf{X}_i = (\mathbf{X}_{11}, \ldots, \mathbf{X}_{kn})$ s realizacemi $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in})$, $i=1,\ldots,k$, potom je

$$P(Y_i = 1 | \mathbf{X}_i = \mathbf{x}_i) = \frac{1}{1 + \mathrm{e}^{-\mathbf{x}_i'\vec{\beta}}}$$

Parametry $\vec{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ odhadujeme metodou maximální věrohodnosti.

Příklad: Zajímá nás vliv několika příznaků (obezita, kouření, alkohol) na vznik onemocnění.

Naplánujeme experiment a provedeme pozorování:

obezita	kouření	alkohol	celkem	nemocných
–	–	–	60	5
+	–	–	17	2
–	+	–	8	1
+	+	–	2	0
–	–	+	187	35
+	–	+	85	13
–	+	+	51	15
+	+	+	23	8

```
> ano.ne  <- c("ano","ne")
> obezita <- gl(2,1,8,ano.ne)
> kouření <- gl(2,2,8,ano.ne)
> alkohol <- gl(2,4,8,ano.ne)
> n.celk  <- c(60,17,8,2,187,85,51,23)
> n.nemoc <- c(5,2,1,0,35,13,15,8)
> hyp.tbl <- cbind(n.nemoc, n.celk-n.nemoc)
```

Model logistické regrese

Příklad: Zajímá nás vliv několika příznaků (obezita, kouření, alkohol) na vznik onemocnění.

obezita	kouření	alkohol	celkem	nemocných
–	–	–	60	5
+	–	–	17	2
–	+	–	8	1
+	+	–	2	0
–	–	+	187	35
+	–	+	85	13
–	+	+	51	15
+	+	+	23	8

```
> nemoc <- glm(hyp.tbl ~ obezita + kouření + alkohol, binomial)
> nemoc

  Call:  glm(formula = hyp.tbl ~ obezita + kouření + alkohol,
  family = binomial("logit"))

  Coefficients:
  (Intercept)    obezitaano    kouřeníano    alkoholano
     -2.37766      -0.06777       0.69531      0.87194

  Degrees of Freedom: 7 Total (i.e. Null);  4 Residual
  Null Deviance:          14.13
  Residual Deviance: 1.618    AIC: 34.54
```

Model logistické regrese

Příklad: Zajímá nás vliv několika příznaků (obezita, kouření, alkohol) na vznik onemocnění.

```
> summary(nemoc)
```

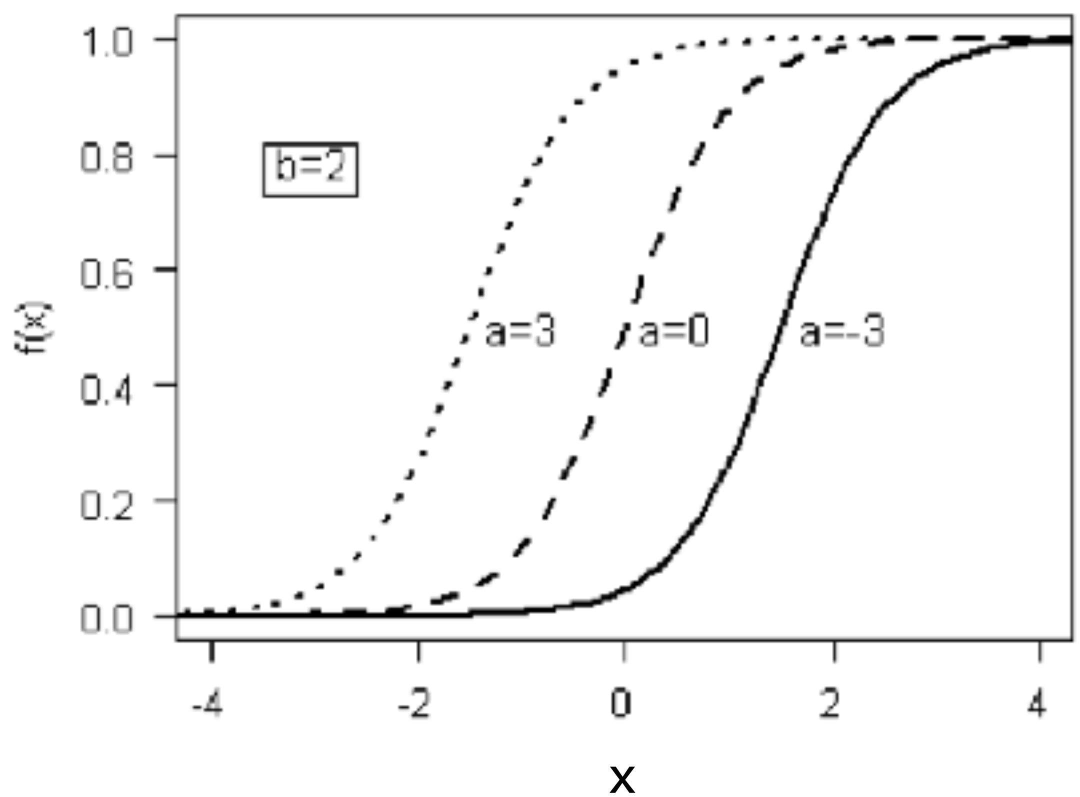
```
Call:
glm(formula = hyp.tbl ~ obezita + kouření + alkohol, family =
binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.37766    0.38018  -6.254    4e-10 ***
obezitaano  -0.06777    0.27812  -0.244   0.8075
kouřeníano   0.69531    0.28509   2.439   0.0147 *
alkoholano   0.87194    0.39757   2.193   0.0283 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14.1259  on 7  degrees of freedom
Residual deviance:  1.6184  on 4  degrees of freedom
AIC: 34.537
```

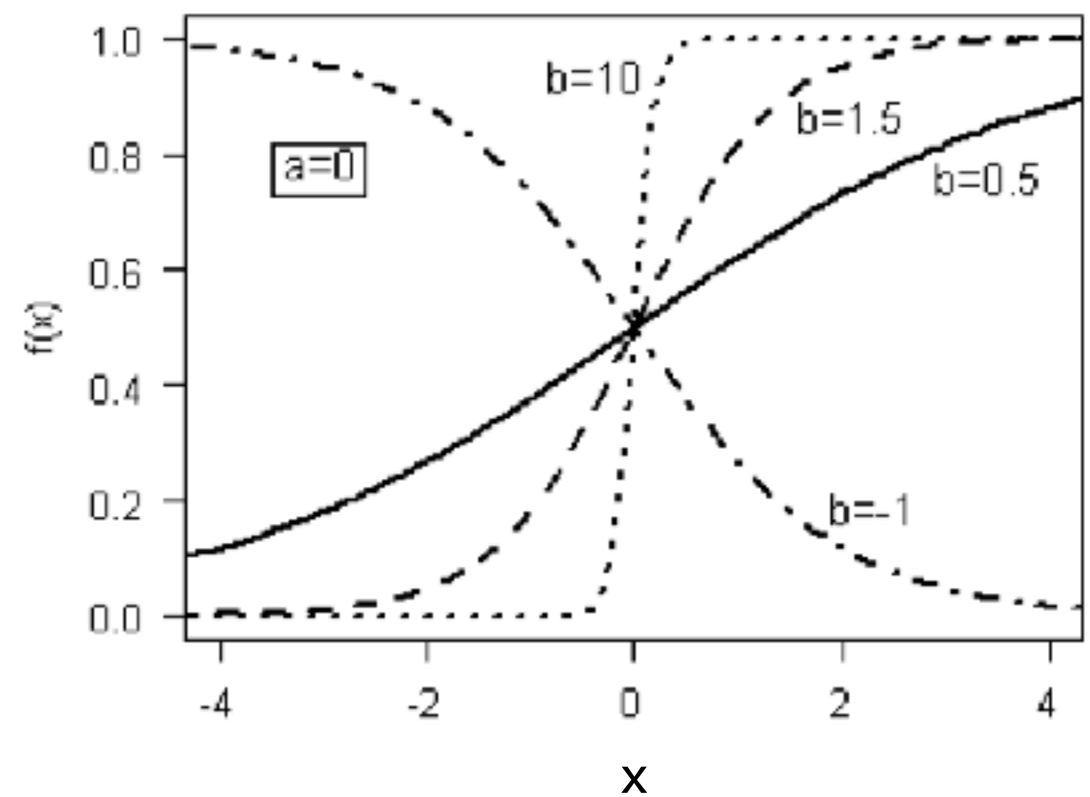
Model logistické regrese

Logistický regresní model s jedním prediktorem

$$P(Y=1|\,X=x)=\frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 x)}}$$

Jak vypadá funkce $f(x;\,a,\,b)=\dfrac{1}{1+\mathrm{e}^{-(a+bx)}}$?





Model logistické regrese

Příklad: Zkoumání závislosti detekce trhliny na její velikosti:

V průběhu experimentu jsou detekovány trhliny v materiálu a je zkoumána závislost pravděpodobnosti této detekce (Y) na velikosti trhliny X v mm (tzv. PoD křivka).

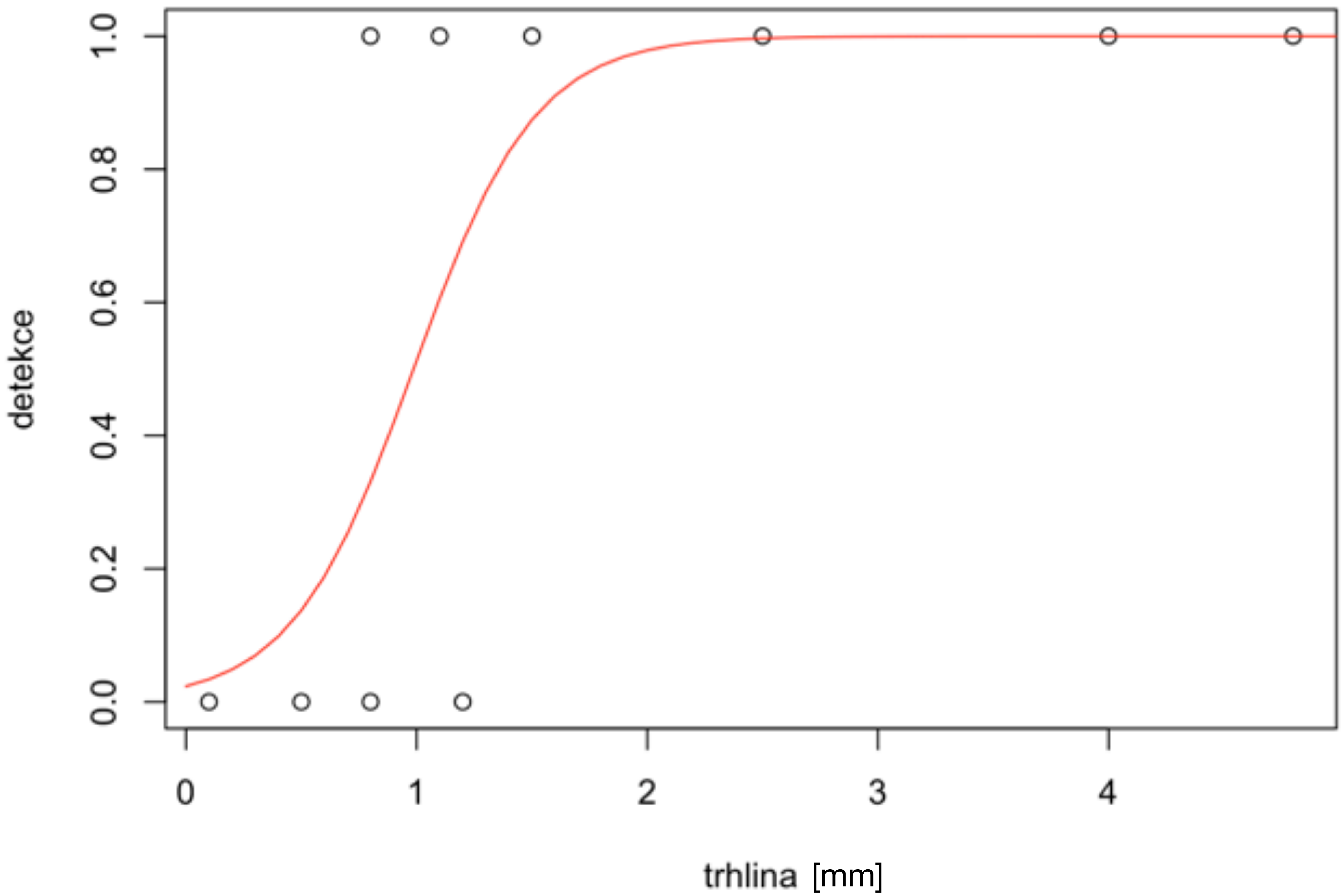
$$P(Y=1|\,X=x)=\frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 x)}}$$

```
> trhlina  <- c(0.1,2.5,0.5,0.8,1.5,1.2,1.1,0.8,4.0,4.8)
> detekce  <- c( 0 , 1 , 0 , 0 , 1 , 0 , 1 , 1 , 1 , 1)
> experiment <- glm(detekce ~ trhlina, binomial)
> summary(experiment)

Call: glm(formula = detekce ~ trhlina, family = binomial)

Coefficients:  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.731      2.943  -1.268    0.205
trhlina          3.781      2.920   1.295    0.195

. . .
```

Model logistické regrese

Příklad: Zkoumání závislosti detekce trhliny na její velikosti:

```
> plot(trhlina,detekce)
> x <- seq(0,5,0.1)
> lines(x,1/(1+exp(-experiment$coefficients[1]-
          experiment$coefficients[2]*x)),col="red")
```



Pravděpodobnost detekce (PoD)

... a to je všechno.
Dál už každý
sám ;)