# LOGISTINĖ REGRESIJA

**Logistic Regression**

# TRUMPAI

**Dvireikšmė (*binary*) logistinė regresija – toks modelis, kai vienam (priklausomam) dvireikšmiui kintamąjam daro įtaką vienas ar keletas (nepriklausomų, aiškinamųjų) kintamųjų.**

**Yra ir daugelio kintamųjų logistinė regresija. Jos nenagrinėsime.**

# PAVYZDŽIAI

- Pagal paciento svorį ir kraujo tyrimus reikia nustatyti tikimybę susirgti diabetu.
- Pagal testų rezultatus siekiama nustatyti, ar reiks kompiuteriui garantinio remonto.
- Aiškinamasi, ar žinant rinkėjo pajamas ir amžių galima numatyti, balsuos jis už kandidatą ar nebalsuos.

# KINTAMIEJI

- Priklausomas kintamasis ***Y*** **–** dvireikšmis (0 arba 1).
- Aiškinamieji kintamieji (***X***) – intervaliniai arba pseudokintamieji.
- Jei ***Y*** įgyja kitokias dvi reikšmes – jis perkoduojamas.
- Vienetai (nuliai) sudaro ne daugiau kaip 80 % ***Y*** stebėjimų.

# Modelis:

$$P(Y=1)=\frac{e^{z(x)}}{1+e^{z(x)}};$$

čia

$$z(x)=\alpha+\beta_1 x_1+\ldots+\beta_k x_k.$$

# Kitas modelio užrašas

$$\ln \frac{P(Y=1)}{P(Y=0)} = z(x);$$

čia

$$z(x) = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k.$$

# Tikslai

- Rasti parametrų $\alpha, \beta_1, ..., \beta_\kappa$ verčius $\hat{\alpha}, \hat{\beta}_1, ..., \hat{\beta}_\kappa$.
- Išsiaiškinti, kaip gerai modelis tinka duomenims.
- Gebėti modelį pritaikyti prognozėms.

# Šiek tiek terminų

Tikimybių santykis

$$\frac{P(Y=1)}{1-P(Y=1)}$$

vadinamas įvykio ***Y* = 1** galimybe (*odds*).

# Galimybių santykis

Koeficiento $\beta_k$ eksponentė $exp\{\beta_k\}$ dar vadinama galimybių santykiu (*odds ratio*).

Galimybių santykis parodo, kaip keičiasi $Y$=1 galimybė, kai $x_k$ padidėja vienetu (kiti x nekinta).

# Logistinės regresijos pavyzdys

Norėdamas sužinoti, ar inkubacinės aplinkos temperatūra turi įtakos vėžliukų lyčiai, Ajovos universiteto profesorius K. Koehler tyrė, kiek kokios lyties vėžliukų išsirito iš skirtingose temperatūrose laikytų vėžlio kiaušinių.

# Duomenys

| Temperatūra | Vėžliukai | Vėžliukės |
|---|---|---|
| 27,2 $C^0$ | 2 | 25 |
| 27,7 $C^0$ | 17 | 7 |
| 28,3 $C^0$ | 26 | 4 |
| 28,4 $C^0$ | 19 | 8 |
| 28,9 $C^0$ | 27 | 1 |

# PASTABOS

- Visos interpretacijos yra tokios pat kaip ir SPSS arba SAS programose.
- Logistinę regresiją galima atlikti įvairiais būdais. Apsiribosime funkcija *glm*.
- Tegul reikšmingumo lygmuo $\alpha = 0{,}05$.

# DUOMENŲ ĮVEDIMAS

```
temp = c(27.2,27.7,28.3,28.4,29.9)
nvyr = c(2,17,26,19,27)
nmot = c(25,7,4,8,1)
nviso = nvyr+nmot
pvyr = nvyr/nviso
response = cbind(nvyr,nmot)
```

# R PROGRAMA LOGISTINEI REGRESIJAI

```
nviso.logit=glm(response~temp,
family=binomial)
summary(nviso.logit)
```

# R PROGRAMOS REZULTATAI

Coefficients:

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -61.3183 | 12.0224 | -5.100 | 3.39e-07 | *** |
| temp | 2.2110 | 0.4309 | 5.132 | 2.87e-07 | *** |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 64.429 on 4 degrees of freedom
Residual deviance: 14.863 on 3 degrees of freedom
AIC: 33.542

# Modelio tinkamumas duomenims

Modelis gerai tinka duomenims, jeigu deviacijos ir laisvės laipsnių santykis mažesnis už vienetą.

Modelis neblogai tinka duomenims, jeigu deviacijos ir laisvės laipsnių santykis nedaug viršija vienetą (pvz., lygus 1,2).

Jeigu santykis daug viršija vienetą, modelis ne itin gerai tinka duomenims.

# TINKAMUMAS DUOMENIMS

Pateikta pati deviacija (14,863) ir jos laisvės laipsniai (3). Suskaičiavę santykį, gauname 14,863/3=4,954. Deviacijos ir laisvės laipsnių santykis rodo blogą modelio tinkamumą duomenims.

Null deviance: 64.429 on 4 degrees of freedom
Residual deviance: 14.863 on 3 degrees of freedom

# PASTABOS

- Modelį reikėtų tobulinti.
- Viena iš galimų didelės deviacijos priežasčių yra labai negausus skirtingų *Temp* reikšmių skaičius. Jų buvo tik penkios.
- Skaičiuojant  laisvės laipsnius, įtakos turi skirtingų aiškinamųjų kintamųjų reikšmių skaičius. Įprastinė GLM formulė (*n-K-1*) netaikoma.
- Toliau aptarsime likusius rezultatus.

# VOLDO TESTAI

- Ieškome nesvarbių x.
- Kiekvienam daugikliui $\beta_m$ tikrinama:

  $\mathbf{H_0: \beta_m = 0,}$

  $\mathbf{H_1: \beta_m \neq 0.}$
- Jei nulinės hipotezės neatmetame – tai kintamasis modelyje galbūt nereikalingas. Reikia patikrinti modelį be šio kintamojo.

# STATISTINĖS IŠVADOS ATSIŽVELGIANT Į *P* REIKŠMĘ

**$H_0$ atmetame (kintamasis modeliui tinka), jei**

$$p < \alpha.$$

**$H_0$ neatmetame (kintamasis „įtartinas“ , jei**

$$p >= \alpha.$$

**Čia $\alpha$ - reikšmingumo lygmuo.**

# Ką daryti su „įtartinais“ kintamaisiais

- **Pakartojame regresijos modelį be tokio kintamojo.**
- **Tiriame klasifikavimo lentelę.**
- **Jei klasifikavimo tikslumas praktiškai nepakito – kintamąjį šaliname.**

Dažniausiai, modelio konstantos Voldo kriterijaus *p* reikšmės net nenagrinėjame, nebent mums labai svarbu, ar konstanta nelygi nuliui.

# VOLDO TESTAI R

Kintamojo *temp* Voldo kriterijaus *p* reikšmė maža (2,87e-07=0,000000287<0,05).

Kintamasis *temp* modelyje reikalingas.

```
Coefficients:
             Estimate Std. Error       z value     Pr(>|z|)
(Intercept) -61.3183    12.0224  -5.100      3.39e-07 ***
temp          2.2110     0.4309    5.132      2.87e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# PARAMETRŲ ĮVERČIAI R

$$\hat{\alpha} = -61{,}32$$

$$\hat{\beta}_1 = 2{,}21$$

$$\hat{z}(x) = -61{,}32 + 2{,}21 \cdot temp.$$

```
Coefficients:
             Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -61.3183    12.0224    -5.100      3.39e-07 ***
temp          2.2110     0.4309     5.132      2.87e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Prognozavimas

Prognozuojama tik tada, kai regresijos modelis duomenims tinka. Taikoma formulė

$$\hat{P}(Y=1)=\frac{e^{\hat{z}(x)}}{1+e^{\hat{z}(x)}}.$$

# Prognozavimo pavyzdys

Kai temperatūra yra 27,5 $C^0$, tai

$$\hat{z}(x) = -61{,}32 + 2{,}21 \cdot 27{,}5 = -0{,}545,$$

$$\hat{P}(Y = 1) = \frac{e^{-0{,}545}}{1 + e^{-0{,}545}} = 0{,}367.$$

# Prognozavimo pavyzdys

Žinome, kad Y=1 atitinka teiginį *išsiris vėžliukas*. Todėl gautąjį rezultatą interpretuojame taip:

esant 27,5 $C^0$ temperatūrai, tikimybės išsiristi vėžliukui įvertis yra 0,367. Tikimybė išsiristi vėžliukei lygi 1- 0,367 = 0,633.

# Galimybės įvertis

Apskaičiuojame

$$\frac{\hat{P}(Y=1)}{1-\hat{P}(Y=1)} = \frac{0{,}367}{0{,}633} = 0{,}58.$$

Darome išvadą, kad esant 27,5 $C^0$ temperatūrai beveik dukart tikėtiniau, kad išsiris vėžliukė nei vėžliukas (tiksliau 100/58 karto tikėtiniau).

# GALIMYBIŲ SANTYKIS

Norint surasti galimybių santykį, reikia įvykdyti tokią R komandą:

```
exp(nviso.logit$coefficients)
```

Gausime:

| (Intercept) | temp |
|---|---|
| 2.343115e-27 | 9.125122e+00 |

# GALIMYBIŲ SANTYKIS

Koeficientas prie kintamojo *temp* buvo įvertintas taip: $\hat{\beta}_1 = 2{,}21$

Daugiklis

9.125122e+00 = 9,125122 = **Exp(2,211)**

rodo, kaip keičiasi galimybių santykis, temperatūrai pakilus vienu laipsniu.

# GALIMYBIŲ SANTYKIS

Galimybių santykį **Exp(2,211)=9,125** interpretuojame taip: temperatūrai padidėjus vienu laipsniu, galimybė išsiristi vėžliukui padidėja 9,125 karto.

*Pastaba. Galimybė – nėra tikimybė, vienetą viršyti gali*.

# Galimybių santykio taikymas

Apskaičiuojame, kaip pasikeis galimybė temperatūrai nuo 27,5 $C^0$ pakilus iki 28,5 $C^0$:

$$0{,}58 \cdot 9{,}125 = 5{,}29.$$

Darome išvadą, kad esant 28,5 $C^0$ **penkis kartus labiau tikėtina**, kad išsiris vėžliukas, o ne vėžliukė.

# BENDRASIS MODELIO TINKAMUMAS

Vertinant bendrąjį modelio tinkamumą duomenims dažniausiai taikomas chi kvadrato kriterijus.

Aptarsime, kokia hipotezė šiuo kriterijumi tikrinama, ir kaip jis realizuotas R programoje.

# $\chi^2$ kriterijus

- **Tikrina hipotezę:**

  **$H_0$: visi $\beta_m = 0$,**

  **$H_1$: ne visi $\beta_m = 0$.**

- **Kitais žodžiais:**

  **$H_0$: Y nepriklauso nuo x,**

  **$H_1$: Y priklauso nuo x.**

- **Tik nežinome, nuo kurių x.**

# Statistinės išvados atsižvelgiant į *p* reikšmę

**$H_0$ atmetame (logistinė regresija galbūt tinka), jei**

$$p < \alpha$$

**$H_0$ neatmetame (logistinė regresija netinka), jei**

$$p >= \alpha$$

**Čia $\alpha$ – reikšmingumo lygmuo.**

# $\chi^2$ kriterijus naudojant R

$\chi^2$ kriterijaus reikšmė gaunama, įvykdžius komandą:

```
nviso.logit$null.deviance-nviso.logit$deviance
```

Gauname $\chi^2$ statistikos reikšmę :

```
49.56555
```

# $\chi^2$ kriterijus naudojant R

$\chi^2$ kriterijaus laisvės laipsnių skaičius gaunamas, įvykdžius komandą:

```
nviso.logit$df.null-nviso.logit$df.residual
```

Gauname, kad laisvės laipsnių yra

```
1
```

# $\chi^2$ kriterijus naudojant R

$\chi^2$ kriterijaus *p* reikšmė gaunama, įvykdžius komandą:

```
dchisq(nviso.logit$null.deviance-
nviso.logit$deviance, nviso.logit$df.null-
nviso.logit$df.residual)
```

Gauta *p* reikšmė 9.77906e-13 yra daug mažesnė už 0,05. Todėl galima teigti, kad $\chi^2$ kriterijaus rodo puikų modelio tinkamumą duomenims.

# PASTABOS:

Visi gautieji rezultatai, išskyrus deviaciją, rodo pakankamai gerą modelio tinkamumą duomenims. Vis dėlto, mažas skirtingų temperatūrų skaičius kelia rimtų įtarimų, kad modelis nėra labai geras.