

# LOGISTINĖ REGRESIJA

**Logistic Regression**

# TRUMPAI

**Dvireikšmė (*binary*) logistinė regresija**  
– toks modelis, kai vienam  
(priklausomam) **dvireikšmiui**  
kintamajam 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

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

# Kitas modelio užrašas

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

čia

$$z(\mathbf{x}) = \alpha + \beta_1 x_1 + \dots + \beta_k x_k.$$

# Tikslai

- Rasti parametrų  $\alpha, \beta_1, \dots, \beta_K$  verčius  $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_K$ .
- 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 išsiritu iš skirtingose temperatūrose laikytų vėžlio kiaušinių.

# Duomenys

Temperatūra Vėžliukai Vėžliukės

27,2 C <sup>0</sup>	2	25
27,7 C <sup>0</sup>	17	7
28,3 C <sup>0</sup>	26	4
28,4 C <sup>0</sup>	19	8
28,9 C <sup>0</sup>	27	1

# LOGISTINĖ REGRESIJA, naudojant **SAS** programą

# DUOMENYS

Duomenis galima įvesti keliais skirtingais būdais. Pateikiame vieną iš jų.

# DUOMENŲ ĮVEDIMAS

```
data vezliukai;  
input temp nvyr nmot;  
nviso=nvyr+nmot;  
datalines;  
27.20 2 25  
27.70 17 7  
28.30 26 4  
28.40 19 8  
29.90 27 1  
run;
```

# SAS PROCEDŪROS

Logistinę regresiją SAS programa galima atlikti dviem būdais:

- Naudojant procedūrą *proc logistic* .
- Nagrinėjant, kaip atskirą GLM atvejį, ir naudojant procedūrą *proc genmod*.

Visi pagrindiniai rezultatai sutaps. Be to, antruoju atveju gausime modelio deviaciją.

# LOGISTINĖ REGRESIJA NAUDOJANT *PROC* *LOGISTIC*

**SAS** programa

# PROCEDŪRA *PROC LOGISTIC*

```
/* logistinė regresija */  
proc logistic data=vezliukai;  
    model nvyr/nviso = temp / rsq  
    lackfit ctable;  
run;
```

# REZULTATAI

Model Information	
Data Set	WORK.VEZLIUKAI
Response Variable (Events)	nvyr
Response Variable (Trials)	nviso
Model	binary logit
Optimization Technique	Fisher's scoring

Modelio reikšmė  $Y=1$  atitiks įvykį – išsiriti vėžliukas.

# $\chi^2$ 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-ų.

# Statistinės išvados atsižvelgiant į $p$ reikšmę

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

$$p < \alpha$$

$H_0$  neatmetame (logistinė regresija netinka), jei

$$p \geq \alpha$$

Čia  $\alpha$  – reikšmingumo lygmuo.

# $\chi^2$ KRITERIJUS

$\chi^2$  kriterijaus reikšmė statistiškai reikšminga. Tai rodo pakankamai gerą modelio tinkamumą duomenims.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	49.5656	1	<.0001
Score	36.9432	1	<.0001
Wald	26.3345	1	<.0001

# HOSMERIO – LEMEŠOU KRITERIJUS

Šis kriterijus – alternatyva anksčiau aptartajam  $\chi^2$  kriterijui. Hosmerio – Lemešou kriterijus aprašytas *Statistika ir jos taikymai.II* (p. 190).

Modelis **nelabai tinka** duomenims, kai  $p$  reikšmė maža ( $p < 0,05$ ).

# HOSMERIO – LEMEŠOU KRITERIJUS

Matome, kad  $p$  reikšmė maža ( $p < 0,05$ ).  
Darome išvadą, kad modelio  
tinkamumas duomenims nėra labai

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
14.9522	3	0.0019

# PASTABA

Taikant  $\chi^2$  ir Hosmerio - Lemešou kriterijus gautos visiškai priešingos išvados apie modelio tinkamumą. Šiuo atveju, ko gero labiau reikėtų tikėti Hosmerio - Lemešou kriterijumi, nes duomenų nėra daug ir  $p$  reikšmė netapo maža vien dėl labai didelės imties.

Taigi modelis nėra labai tinkamas duomenims. Vis dėlto nekeisdami modelio aptarsime ir kitus rodiklius.

# VOLDO TESTAI

- Ieškome nesvarbių x-ų.
- Kiekvienam daugikliui  $\beta_m$  tikrinama:  
 $H_0: \beta_m = 0,$   
 $H_1: \beta_m \neq 0.$
- 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

$$p \geq \alpha.$$

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 SAS

Kintamojo *temp* Voldo kriterijaus  $p$  reikšmė maža. Kintamasis modelyje reikalingas.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-61.3183	12.0224	26.0134	<.0001
temp	1	2.2110	0.4309	26.3345	<.0001

# PARAMETRŲ ĮVERČIAI SAS

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

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

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

## Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-61.3183	12.0224	26.0134	<.0001
temp	1	2.2110	0.4309	26.3345	<.0001

# PROGNOZAVIMAS

- Konkretiems  $x_m$  galima apskaičiuoti

$$\hat{z}(x) = \hat{\alpha} + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$$

ir rasti  $P(Y = 1)$  įvertį.

# 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<sup>0</sup>, 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  $27,5\text{ C}^0$  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 išvada, kad esant 27,5 C<sup>0</sup> temperatūrai beveik dukart tikėtiniau, kad išsiris vėžliukė nei vėžliukas (tiksliau 100/58 karto tikėtiniau).

# Galimybių santykis

Daugiklis  $\text{Exp}(2,211) = 9,125$  rodo, kaip keičiasi galimybių santykis, temperatūrai pakilus vienu laipsniu.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
temp	9.125	3.922	21.232

# Galimybių santykis

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

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

# Galimybių santykio taikymas

Apskaičiuojame, kaip pasikeis galimybė temperatūrai nuo 27,5 C<sup>0</sup> pakilus iki 28,5 C<sup>0</sup>:

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

Darome išvadą, kad esant 28,5 C<sup>0</sup> **penkis kartus labiau tikėtina**, kad išsiris vėžliukas, o ne vėžliukė.

# DETERMINACIJOS KOEFICIENTAI

- Jų yra net keli. Dažniausiai naudojami Kokso-Snelo arba Nagelkerkės determinacijos koeficientai.
- Kuo  $R^2$  didesnis (arčiau vieneto), tuo modelis geresnis.
- Mažas  $R^2$  rodo, kad logistinės regresijos modelis duomenims nelabai tinka.
- SAS programoje Kokso-Snelo koeficientas vadinamas tiesiog  $R^2$ .

# DETERMINACIJOS KOEFICIENTAI

SAS programoje Nagelkerkės determinacijos koeficientas vadinamas *Max-rescaled R-Square*.

Nagelkerkės determinacijos koeficientas lygus 0,4248. Tai – vidutinis didumas, rodantis neblogą modelio tinkamumą duomenims.

<b>R-Square</b>	0.3054	<b>Max-rescaled R-Square</b>	0.4248
-----------------	--------	------------------------------	--------

# KITI RODIKLIAI

Sudaromos visos įmanomos duomenų  $Y=1$  ir  $Y=0$  poros. (*Pairs*). Galime palyginti ir apskaičiuotuosius tikimybių įverčius. Jeigu  $P(Y=1)$  įvertis didesnis už  $P(Y=0)$  įvertį, tai turime *suderintą stebėjimų porą* (*concordant pair*). Priešingu atveju – pora *nesuderinta* (*disconcordant pair*). Kai tikimybės lygios, turime lygiavertę porą (*tied pair*). Kuo didesnis suderintų porų procentas, tuo modelis geriau tinka duomenims.

# SOMERSO D

Porų suderinamumu grindžiamas ir Somerso D koeficientas .

Somerso D gali įgyti bet kokią reikšmę iš intervalo  $[-1, 1]$ .

Kai  $D=-1$ , turime labai blogai duomenims tinkantį modelį.

Kai  $D=1$ , modelis duomenims tinka idealiai.

# PORŲ SUDERINAMUMAS

Suderintų porų yra 76,4 %. Somerso  $D=0,639$ .  
Abudu rodikliai rodo pakankamai gerą  
modelio tikimą duomenims.

<b>Association of Predicted Probabilities and Observed Responses</b>			
<b>Percent Concordant</b>	76.4	<b>Somers' D</b>	0.639
<b>Percent Discordant</b>	12.6	<b>Gamma</b>	0.718
<b>Percent Tied</b>	11.0	<b>Tau-a</b>	0.285
<b>Pairs</b>	4095	<b>c</b>	0.819

# RODIKLIAI

**Primename, į ką reikia atsižvelgti, taikant logistinę regresiją:**

- į išvadas, gautas pritaikius  $\chi^2$  ir Hosmerio – Lemeshou kriterijus;
- į Voldo kriterijų, nustatant „įtartinus“ aiškinamuosius kintamuosius;
- į determinacijos koeficientus;
- į Somerso D.

## PASTABOS:

1. Visi gautieji rezultatai, išskyrus Hosmerio – Lemešou kriterijų, rodo pakankamai gerą modelio tikimą duomenims. Vis dėlto, mažas skirtingų temperatūrų skaičius kelia rimtų įtarimų, kad modelis nėra labai geras.
2. Taikydami *prod logistic* parinktį negavome labai svarbaus rodiklio – deviacijos. Ją galima rasti, nagrinėjant logistinę regresiją, kaip dalinį apibendrintų tiesinių modelių atvejį.

# LOGISTINĖ REGRESIJA, KAIP DALINIS *GLM* ATVEJIS

**SAS** programa

# PROCEDŪRA *proc genmod*

```
/* GLM */  
proc genmod data=vezliukai;  
model nvyr/nviso = temp / dist = bin  
link = logit lrci obstat residuals;  
run;
```

# REZULTATAI

Reikšmė  $Y=1$  atitinka įvykį *išsiriti vėžliukas*. Logistinės regresijos modelis bus sudarytas tikimybei, kad išsiris vėžliukas.

Model Information	
Data Set	WORK.VEZLIUKAI
Distribution	Binomial
Link Function	Logit
Response Variable (Events)	nvyr
Response Variable (Trials)	nviso

# REZULTATAI

Buvo 5 skirtingos temperatūros reikšmės (jų skaičių atsižvelgiama, skaičiuojant deviacijos laisvės laipsnius), iš 136 kiaušinių 91 kartą išsirito vėžliukas.

<b>Number of Observations Read</b>	5
<b>Number of Observations Used</b>	5
<b>Number of Events</b>	91
<b>Number of Trials</b>	136

# Modelio tinkamumas duomenims

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

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

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

# TIKIMAS DUOMENIMS

Modelis ne itin gerai tinka duomenims, nes deviacijos ir laisvės laipsnių santykis daug didesnis už vieneta (lygus 4,954).

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	3	14.8629	4.9543
Scaled Deviance	3	14.8629	4.9543
Pearson Chi-Square	3	14.9522	4.9841
Scaled Pearson X2	3	14.9522	4.9841
Log Likelihood		-61.5503	
Full Log Likelihood		-14.7712	
AIC (smaller is better)		33.5425	

# 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.

# PARAMETRŲ ĮVERČIAI IR VOLDO KRITERIJUS

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-61.3183	12.0224	-86.8303	-39.7213	26.01	<.0001
temp	1	2.2110	0.4309	1.4377	3.1257	26.33	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

# PARAMETRŲ ĮVERČIAI IR VOLDO KRITERIJUS

Parametrų įverčiai ir jų statistinis reikšmingumas nagrinėjama visiškai analogiškai, kaip ir naudojant procedūrą *proc logistic*.

# KITI RODIKLIAI

Tarp rezultatų yra informaciniai indeksai (Akaičės AICC ir BIC). Juos galima naudoti, kai turime daug aiškinamųjų kintamųjų ir norime dalies jų atsisakyti.

Sudaromas naujas modelis ir lyginamas su ankstesniuoju. Geresnis tas modelis, kurio informaciniai indeksai mažesni.

# METODINĖS PASTABOS

Logistinę regresiją reikėtų daryti naudojant procedūrą *proc logistic*.

Po to pakartoti tyrimą naudojant procedūrą *proc genmod* ir patikrinti, ar deviacijos ir jos laisvės laipsnių santykis daug neviršija vieneto.