Logistická regrese

POLb1139

Peter Spáč

Která regrese?

- Výběr závisí na podobě závislé proměnné

- Lineární (OLS) regrese
  - Kardinální proměnná

- Logistická regrese
  - **Binární ZP (0/1) – binární logistická regrese**
  - Nominální ZP s > 2 hodnotami (0/1/2/3) – multinomiální logistická regrese

- Nejedná se o konečný výčet

Chcete získat titul bc. politologie?

Má smysl…

…účastnit se přednášek?

…číst povinnou literaturu?

…učit se na zkoušky NEJEN z prezentací a výpisků?

…odevzdávat seminární práce dříve než 20 minut před termínem?

…spoléhat se na kofein a energetické nápoje během zk období?

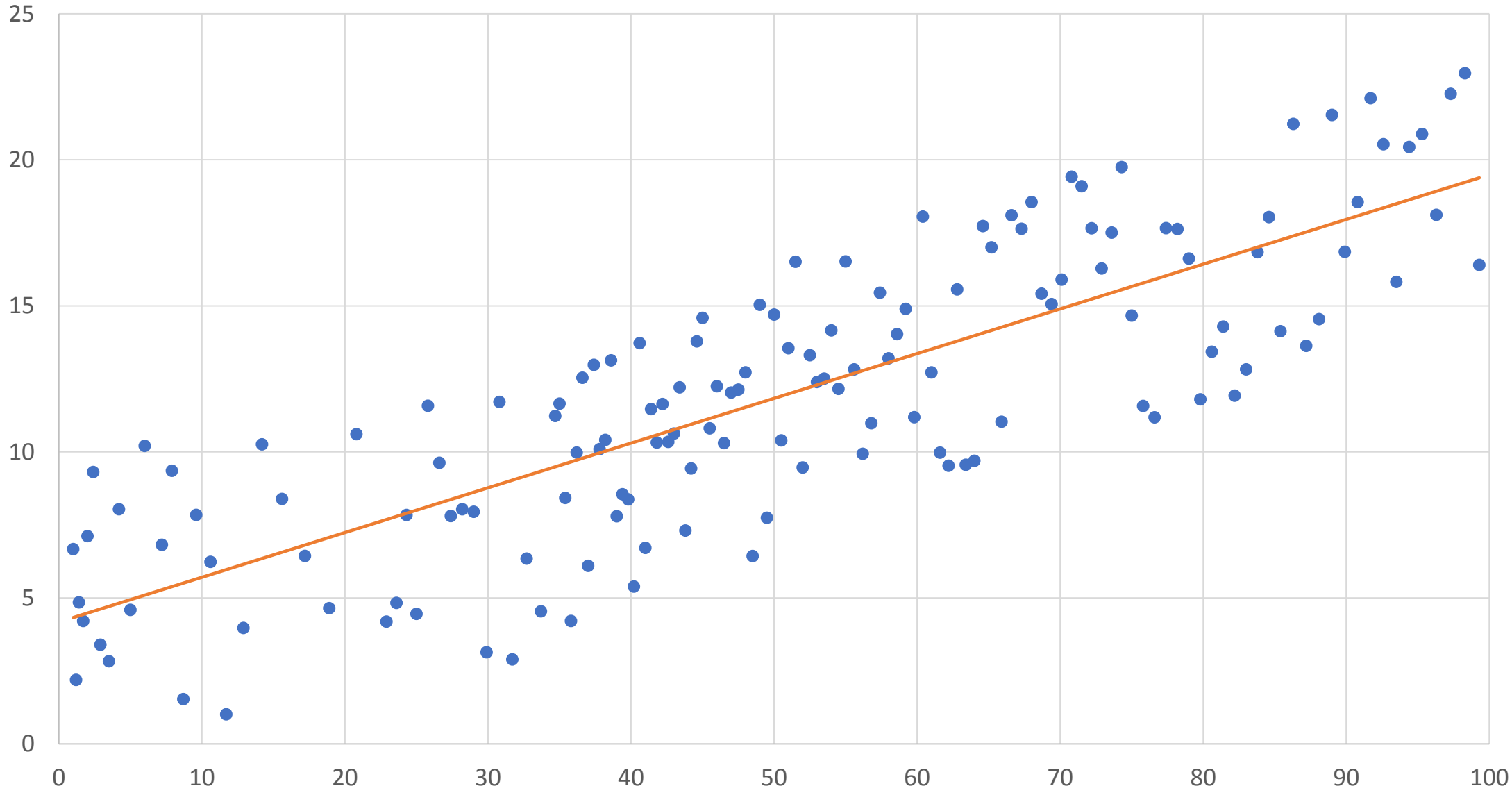
…zdravit své vyučující na chodbách fakulty?

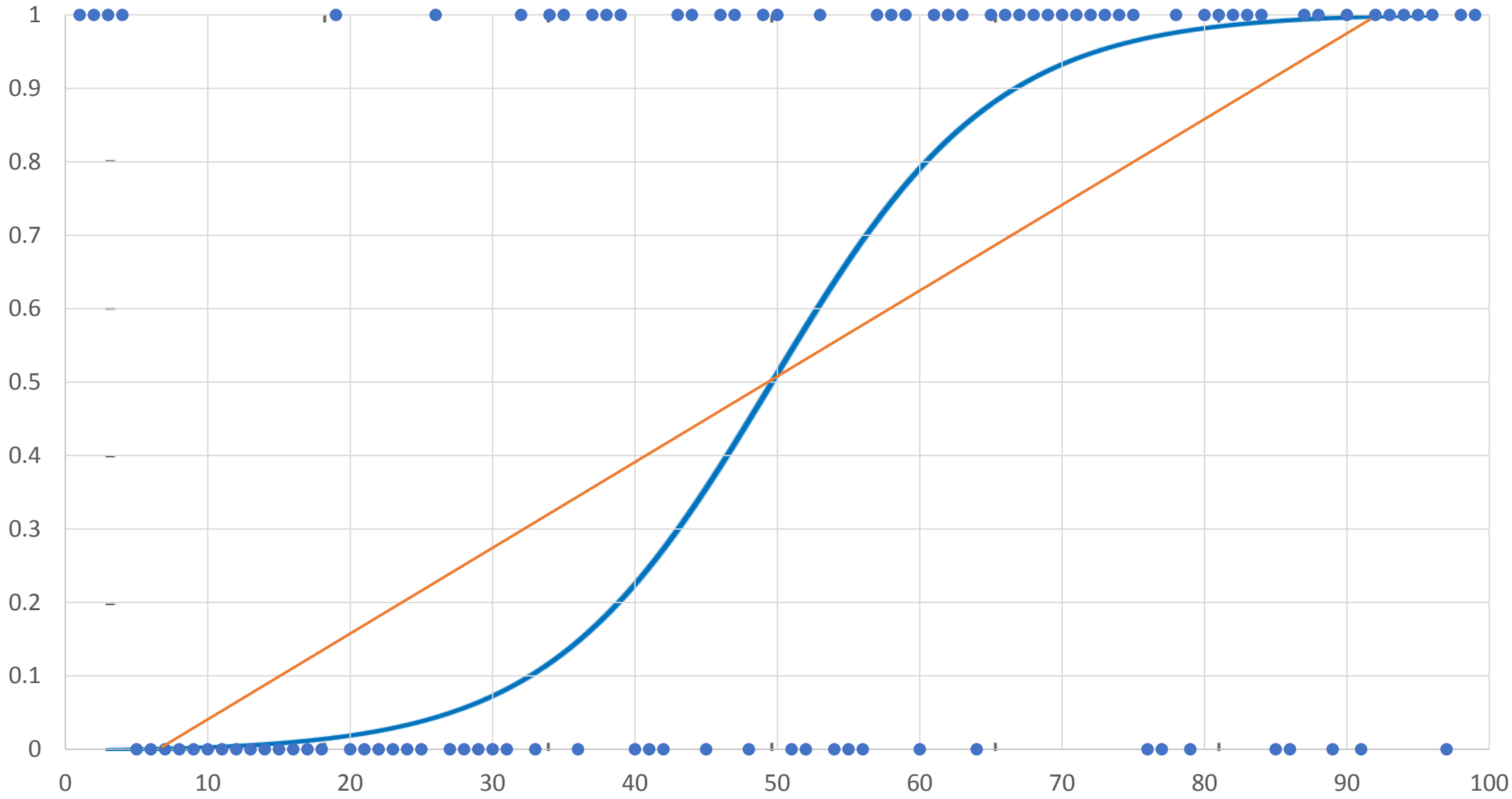
Binární logistická regrese - požadavky

- Závislá proměnná (outcome):
  - Přesně jedna proměnná, binární (0/1)

- Nezávislé proměnné (predictors):
  - Jedna nebo více proměnných, přípustné jsou všechny druhy (možnost rekódování)

- Některé další požadavky:
  - Nezávislost pozorování
  - Absence multikolinearity mezi NP
  - Normální distribuce chyb


25
20
15
10
5
0
0
10
20
30
40
50
60
70
80
90
100


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
100

Výstupy **lineární** regrese

- Parametry:
  - Konstanta
  - Efekt každé NP

- $y = b_0 + b_1*x + b_2*y + b_3*z + \ldots$

- ***y*** – odhadovaná hodnota ZP
- $\boldsymbol{b_0}$ - konstanta
- $\boldsymbol{b_1}$, $\boldsymbol{b_2}$, $\boldsymbol{b_3}$ – nestandardizované beta koeficienty za každou NP
- **x, y, z** – hodnoty NP

Lineární vs. logistická regrese

Odhad hodnoty Y

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_{ni}$$

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_{ni})}}$$

Odhad pravděpodobnosti, že nastane jev Y

Výstupy **logistické** regrese

- Vhodnost modelu (fit)

- Parametry:
  - Konstanta
  - Efekt každé NP

- Pro jednodušší vyjádření můžeme výstupy převést na šance (a ještě lépe na pravděpodobnosti)

Vhodnost modelu

- AIC, BIC:
  - Akaike Information Criterion, Bayesian Information Criterion
  - Ukazují, jak model pasuje na analyzovaná data
  - Vyšší hodnoty ukazují na slabší sílu modelu a naopak

- R^2 (Index determinace):
  - Podobný význam jako u lineární regrese, nejedná se ale o ekvivalent
  - Vyšší hodnota signalizuje vhodnější model
  - Více druhů – JASP počítá čtyři

Informace o efektech NP

- Regresní koeficient b:
  - JASP uvádí jako *Estimate*
  - Jak se při zvýšení hodnoty nezávislé proměnné o jednotku změní **logaritmus** hodnoty závislé proměnné
  - Komplikovanější interpretace

- Poměr šancí (Odds Ratio):
  - Ukazuje, jak se se zvýšením nezávislé proměnné o jednotku mění **šance** na to, že nastane konkrétní výstup v závislé proměnné
  - = exp(Estimate)

Pravděpodobnost

- Nemýlit si šance a pravděpodobnosti!
- V logistické regresi vyjadřuje, jaká je při stanovených hodnotách NP pravděpodobnost, že pro daný zkoumaný případ nastane v ZP jev kódovaný hodnotou 1
- Od 0 po 1, resp. od 0 po 100 %
- Např. muž s VŠ vzděláním ve věku 45 let má 65 % pravděpodobnost, že se zúčastní parlamentních voleb
- P = Odds / (1 + Odds)

Příklad

- Závisely vyhlídky na záchranu pasažérů na Titanicu na jejich věku, pohlaví a třídě?

- H1: Ženy na Titanicu měly vyšší pravděpodobnost záchrany
- H2: S rostoucím věkem pasažérů na Titanicu klesala pravděpodobnost jejich záchrany
- H3: Pasažéři ve vyšších třídách měli vyšší pravděpodobnost záchrany
- Doplníte H0 k H1 / H2 / H3?

- Závislá proměnná:
  - Záchrana – 0/1

- Nezávislé proměnné:
  - Pohlaví – 0/1 (0 = žena, 1 = muž)
  - Věk – věk pasažéra v letech
  - Třída – 1/2/3

JASP

- Regression > Logistic Regression

- Proměnné:
  - Dependent Variable – Survived
  - Covariates – Age
  - Factors – Sex, Pclass

Model Summary - Survived

Model	Deviance	AIC	BIC	df	X^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2	Cox & Snell R^2
H_0	1182.770	1184.770	1189.558	886						
H_1	801.572	811.572	835.511	882	381.198	< .001	0.322	0.474	0.392	0.349

- AIC / BIC ukazují, že model snižuje podíl toho, co nevíme = přispívá k lepšímu poznání variability ZP

- R^2 indikuje podobný závěr

Coefficients

				Wald Test		
	Estimate	Standard Error	z	Wald Statistic	df	p
(Intercept)	3.636	0.371	9.812	96.284	1	< .001
Age	−0.034	0.007	−4.789	22.933	1	< .001
Sex (male)	−2.589	0.187	−13.842	191.594	1	< .001
Pclass (2)	−1.199	0.262	−4.584	21.012	1	< .001
Pclass (3)	−2.455	0.253	−9.697	94.034	1	< .001

Note. Survived level '1' coded as class 1.

- Estimate (regresní koeficient b):
  - Jak se při zvýšení hodnoty nezávislé proměnné o jednotku změní **logaritmus** hodnoty závislé proměnné
  - Věk (- 0,034) – s každým zvýšením věku pasažéra o jeden rok se logaritmus hodnoty ZP sníží o 0,034

Coefficients

				Wald Test		
	Estimate	Standard Error	z	Wald Statistic	df	p
(Intercept)	3.636	0.371	9.812	96.284	1	< .001
Age	−0.034	0.007	−4.789	22.933	1	< .001
Sex (male)	−2.589	0.187	−13.842	191.594	1	< .001
Pclass (2)	−1.199	0.262	−4.584	21.012	1	< .001
Pclass (3)	−2.455	0.253	−9.697	94.034	1	< .001

Note. Survived level '1' coded as class 1.

- Estimate (regresní koeficient b):
  - Jak se při zvýšení hodnoty nezávislé proměnné o jednotku změní **logaritmus** hodnoty závislé proměnné
  - Pohlaví – binární proměnná → ref. kategorie = žena
  - Pohlaví (-2,589) – logaritmus hodnoty ZP je pro muže o 2,589 nižší než pro ženy

Coefficients

				Wald Test		
	Estimate	Standard Error	z	Wald Statistic	df	p
(Intercept)	3.636	0.371	9.812	96.284	1	< .001
Age	−0.034	0.007	−4.789	22.933	1	< .001
Sex (male)	−2.589	0.187	−13.842	191.594	1	< .001
Pclass (2)	−1.199	0.262	−4.584	21.012	1	< .001
Pclass (3)	−2.455	0.253	−9.697	94.034	1	< .001

Note. Survived level '1' coded as class 1.

- Estimate (regresní koeficient b):
  - Jak se při zvýšení hodnoty nezávislé proměnné o jednotku změní **logaritmus** hodnoty závislé proměnné
  - Třída – binární (dummy) proměnné → ref. kategorie = první třída
  - Druhá třída (-1,199) – logaritmus hodnoty ZP je pro pasažéry ve druhé třídě o 1,199 nižší než pro pasažéry první třídy
  - Třetí třída (-2,455) – logaritmus hodnoty ZP je pro pasažéry ve třetí třídě o 2,455 nižší než pro pasažéry první třídy

Coefficients

				Wald Test		
	Estimate	Standard Error	z	Wald Statistic	df	p
(Intercept)	3.636	0.371	9.812	96.284	1	< .001
Age	−0.034	0.007	−4.789	22.933	1	< .001
Sex (male)	−2.589	0.187	−13.842	191.594	1	< .001
Pclass (2)	−1.199	0.262	−4.584	21.012	1	< .001
Pclass (3)	−2.455	0.253	−9.697	94.034	1	< .001

Note. Survived level '1' coded as class 1.

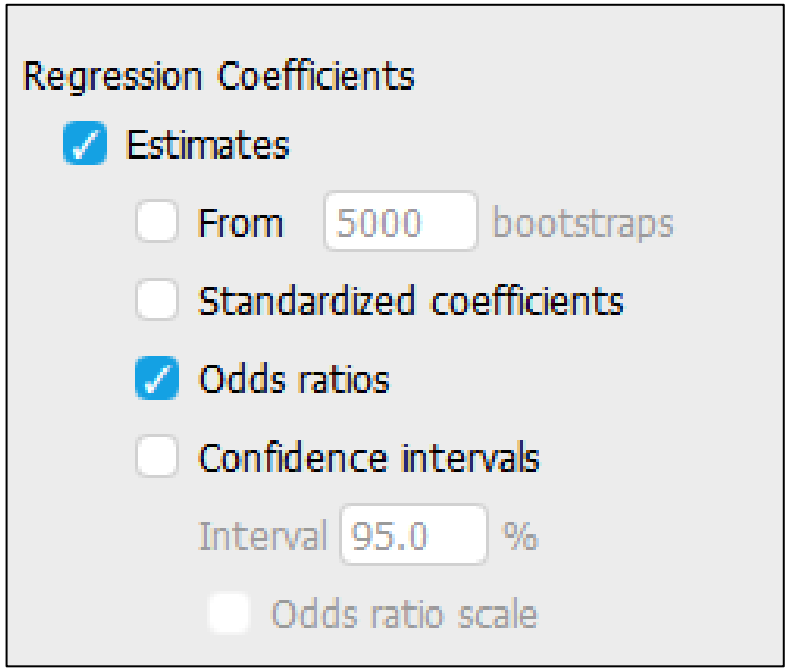
- Estimate (regresní koeficient b):
  - V překladu lepší vyhlídky na záchranu měli mladší lidé, ženy a cestující v první třídě
  - Jak tuto informaci podat publiku? V podobě logaritmu?

- Kdyby nás zajímala signifikantnost, tak všechny efekty jsou statisticky signifikantní

Poměr šancí (Odds ratio)

- Ukazuje, jak se se zvýšením nezávislé proměnné o jednotku mění šance na to, že nastane konkrétní výstup (vyšší hodnota) v závislé proměnné

- Jednodušší interpretace
  - OR = 1 → žádný efekt NP
  - OR > 1 → nárůst NP šance zvyšuje
  - OR < 1 → nárůst NP šance snižuje



Coefficients

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	3.636	0.371	37.932	9.812	96.284	1	< .001
Age	−0.034	0.007	0.966	−4.789	22.933	1	< .001
Sex (male)	−2.589	0.187	0.075	−13.842	191.594	1	< .001
Pclass (2)	−1.199	0.262	0.301	−4.584	21.012	1	< .001
Pclass (3)	−2.455	0.253	0.086	−9.697	94.034	1	< .001

Note. Survived level '1' coded as class 1.

Coefficients

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	3.636	0.371	37.932	9.812	96.284	1	< .001
Age	−0.034	0.007	0.966	−4.789	22.933	1	< .001
Sex (male)	−2.589	0.187	0.075	−13.842	191.594	1	< .001
Pclass (2)	−1.199	0.262	0.301	−4.584	21.012	1	< .001
Pclass (3)	−2.455	0.253	0.086	−9.697	94.034	1	< .001

Note. Survived level '1' coded as class 1.

- OR a konstanta (Intercept):
  - Žena s věkem 0 let (< 1 rok) cestující v první třídě
  - Estimate – logaritmus ZP = 3,636
  - Tato osoba by dle modelu měla šanci na záchranu 38 k jedné

Coefficients

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	3.636	0.371	37.932	9.812	96.284	1	< .001
Age	−0.034	0.007	0.966	−4.789	22.933	1	< .001
Sex (male)	−2.589	0.187	0.075	−13.842	191.594	1	< .001
Pclass (2)	−1.199	0.262	0.301	−4.584	21.012	1	< .001
Pclass (3)	−2.455	0.253	0.086	−9.697	94.034	1	< .001

Note. Survived level '1' coded as class 1.

- OR a efekty NP:
  - Věk – OR = 0,966
  - Každé zvýšení věku osoby o jeden rok snižuje šanci na záchranu o 3,4 % (matematicky se šance násobí číslem 0,966)
  - Např. nárůst věku o tři roky snižuje šanci na záchranu o téměř 10 procent (0,966 * 0,966 * 0,966 = 0,901 → 90,1 %)

Coefficients

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	3.636	0.371	37.932	9.812	96.284	1	< .001
Age	−0.034	0.007	0.966	−4.789	22.933	1	< .001
Sex (male)	−2.589	0.187	0.075	−13.842	191.594	1	< .001
Pclass (2)	−1.199	0.262	0.301	−4.584	21.012	1	< .001
Pclass (3)	−2.455	0.253	0.086	−9.697	94.034	1	< .001

Note. Survived level '1' coded as class 1.

- OR a efekty NP:
  - Pohlaví – OR = 0,075
  - Ve srovnání se ženami měli muži na Titanicu o 92,5 % nižší šanci na záchranu

Coefficients

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	3.636	0.371	37.932	9.812	96.284	1	< .001
Age	−0.034	0.007	0.966	−4.789	22.933	1	< .001
Sex (male)	−2.589	0.187	0.075	−13.842	191.594	1	< .001
Pclass (2)	−1.199	0.262	0.301	−4.584	21.012	1	< .001
Pclass (3)	−2.455	0.253	0.086	−9.697	94.034	1	< .001

Note. Survived level '1' coded as class 1.

- OR a efekty NP:
  - Druhá třída – OR = 0,301
  - Třetí třída – OR = 0,086
  - Ve srovnání s cestujícími v první třídě měli pasažéři ve druhé třídě na Titanicu o téměř 70 % nižší šanci na záchranu
  - Ve srovnání s cestujícími v první třídě měli pasažéři ve třetí třídě na Titanicu o více než 91 % nižší šanci na záchranu

Pravděpodobnost (Probability)

- V logistické regresi vyjadřuje, jaká je při stanovených hodnotách NP pravděpodobnost, že pro daný zkoumaný případ nastane v ZP jev kódovaný hodnotou 1 (cestující na Titanicu se zachránil)

- Optimální způsob vyjádření efektů NP
  - Jednoduchý výpočet
  - Nejpřístupnější forma pro publikum
  - Umožňuje odchýlit se od „technického“ jazyka při interpretaci výsledků

- JASP umožňuje názornou vizualizaci pravděpodobností
  - Nanejvýš vhodné pro kardinální NP

Spočítání pravděpodobnosti

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ \ldots \ + b_n X_{ni}$$

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_{ni})}}$$

- Chceme spočítat pravděpodobnost záchrany cestujícího na Titanicu s vybranými vlastnosti – muž, 15 let, druhá třída
  - Age = 15
  - Sex = 1
  - Pclass (2) = 1
  - Pclass (3) = 0

Coefficients

	Estimate	Standard Error	Odds Ratio	z	Wald Test		
					Wald Statistic	df	p
(Intercept)	3.636	0.371	37.932	9.812	96.284	1	< .001
Age	−0.034	0.007	0.966	−4.789	22.933	1	< .001
Sex (male)	−2.589	0.187	0.075	−13.842	191.594	1	< .001
Pclass (2)	−1.199	0.262	0.301	−4.584	21.012	1	< .001
Pclass (3)	−2.455	0.253	0.086	−9.697	94.034	1	< .001

Note. Survived level '1' coded as class 1.

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i} + b_4 X_{4i})}}$$

- Chceme spočítat pravděpodobnost záchrany cestujícího na Titanicu s vybranými vlastnosti – muž, 15 let, druhá třída
  - Age = 15
  - Sex = 1
  - Pclass (2) = 1
  - Pclass (3) = 0

Coefficients

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	3.636	0.371	37.932	9.812	96.284	1	< .001
Age	−0.034	0.007	0.966	−4.789	22.933	1	< .001
Sex (male)	−2.589	0.187	0.075	−13.842	191.594	1	< .001
Pclass (2)	−1.199	0.262	0.301	−4.584	21.012	1	< .001
Pclass (3)	−2.455	0.253	0.086	−9.697	94.034	1	< .001

Note. Survived level '1' coded as class 1.

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4)}}$$

$$b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i} + b_4 X_{4i}$$

$b_0 = 3{,}636$

$b_1 = -0{,}034$ $\quad X_1 = 15$

$b_2 = -2{,}589$ $\quad X_2 = 1$

$b_3 = -1{,}199$ $\quad X_3 = 1$

$b_4 = -2{,}455$ $\quad X_4 = 0$

$= 3{,}636 - 0{,}034*15 - 2{,}589*1 - 1{,}199*1 - 2{,}455*0$

$= 3{,}636 - 0{,}51 - 2{,}589 - 1{,}199 - 0$

$= \mathbf{-0{,}662}$

<u>Spočítáme šance (Odds):</u>

exp (-0,662) = **0,516**

→ daný cestující má šanci na záchranu 0,516 k jedné

<u>Spočítáme pravděpodobnost:</u>

P = Odds / (1 + Odds) = 0,516 / (1 + 0,516) = 0,516 / 1,516

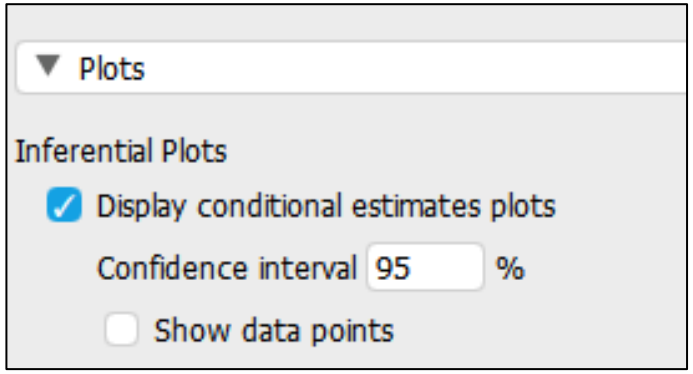
P = 34 %

→ daný cestující má 34 % pravděpodobnost, že přežije

Vizualizace pravděpodobností

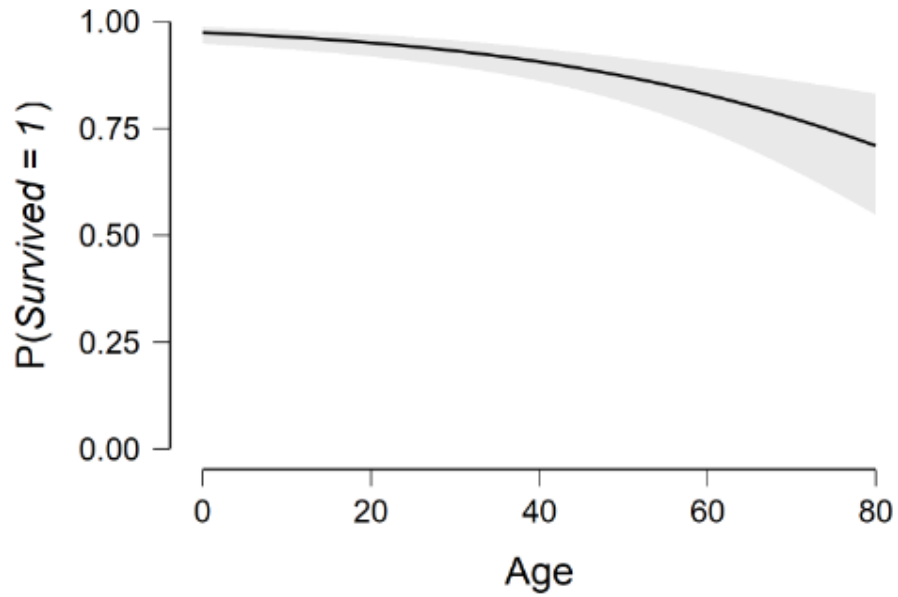
- Vizualizace prozradí o výsledcích víc než tisíc slov
  - Vhodné pro všechny druhy NP
  - Mimořádně vhodné pro zobrazení efektu kardinálních NP

- Na rozdíl od lineární regrese, JASP při logistické regresi příslušné grafy kreslí

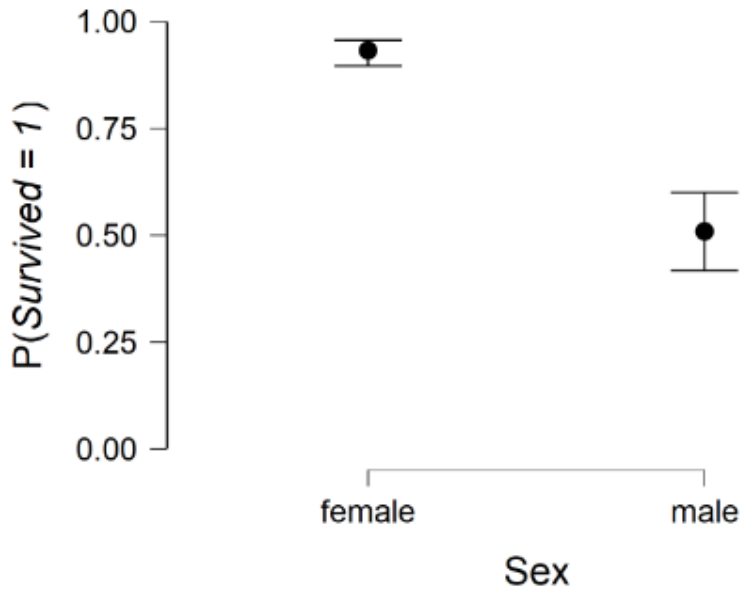


Estimates plots

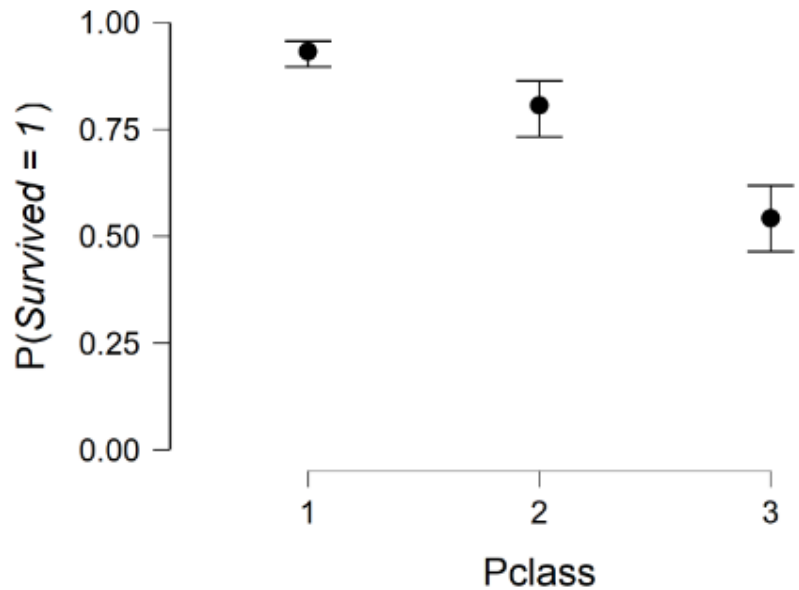
Age



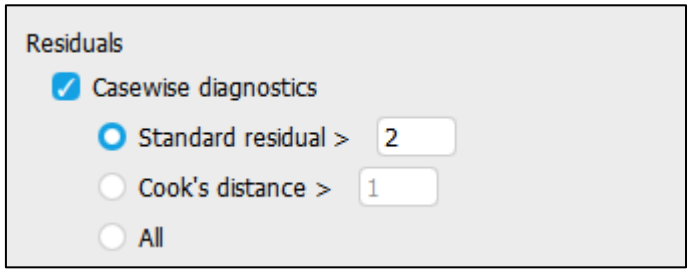
Sex



Pclass



Kontrola předpokladů



- Kontrola reziduí
- Případy s nadměrným vlivem na model (Cook’s distance)
- Multikolinearita
  - Postup jako u lineární regrese
  - VIF nad 5 (10) nebo Tolerance nižší než 0,2 (0,1) indikují problém
  - JASP nemá samostatné testování pro logistickou regresi