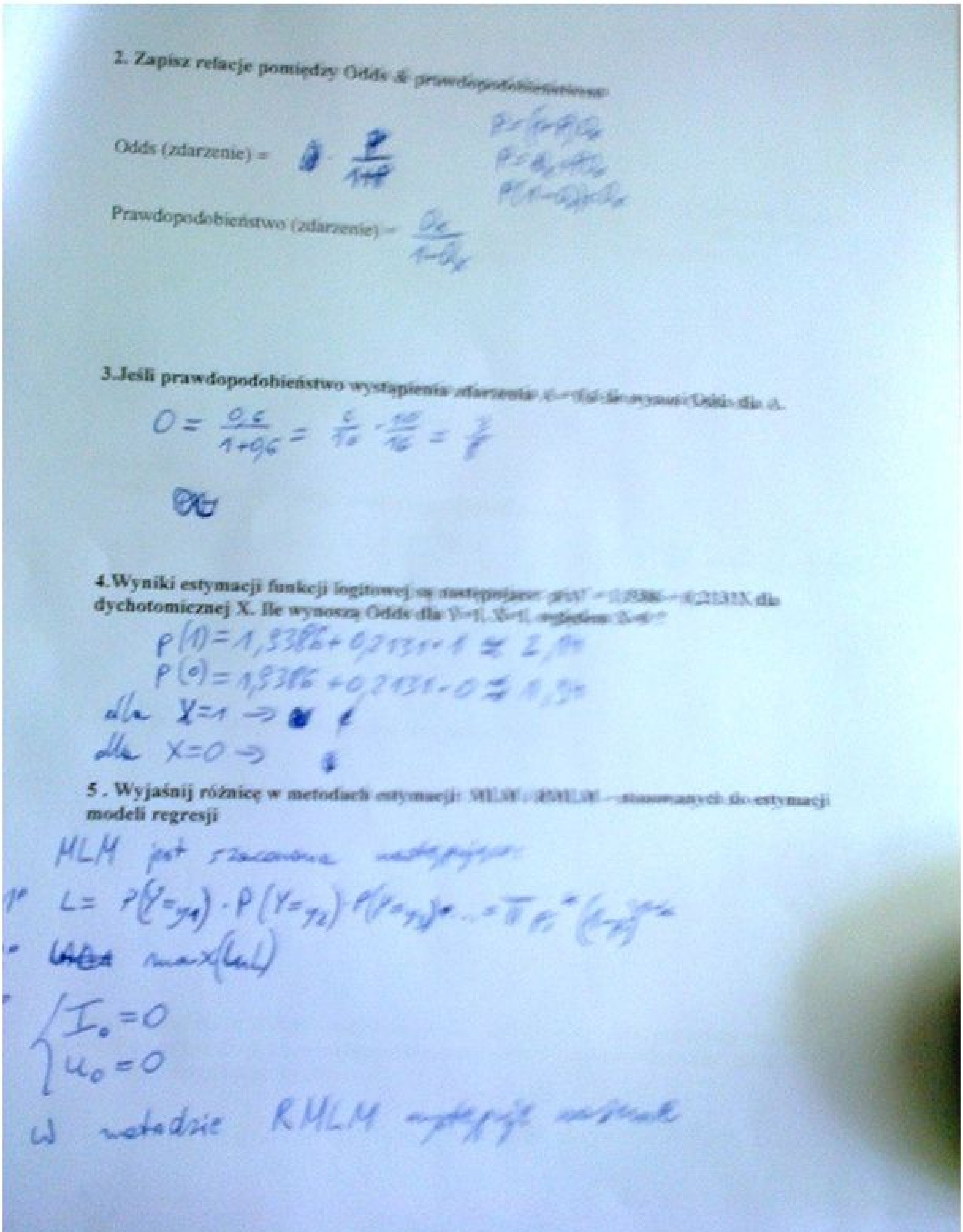
item	zapis postaci	właściwości
Funkcja logistyczna	$f(z) = \frac{1}{1+e^{-z}}$	$\lim_{z \to -nieskon} = 0$, $\lim_{z \to nieskon} = 1$
odds	$Odds = \frac{p}{1-p}$	Iloraz prawdopodobieństwa zajścia zdarzenia i nie zajścia, z (0,1) do (0,nieskończoność)
logit	ln(p/(1-p))	transformuje zmienna z zakresu 0-1na zakres -oo,+oo
ryzyko względne (RR)	$\frac{P(Y=1 \vert X1=1, X2, X3...-const)}{P(Y=1 \vert X1=0, X2, X3....-const)}$	Ile razy prawdopodobieństwo zajścia zdarzenia jest większe w grupie x1=1 (eksperymentalna) niż w grupie x1=0 (kontrolna)
iloraz odds (OR)	$OR = \frac{Odds\ dla\ X1}{Odds\ dla\ X0} = \frac{\frac{P(X1)}{1-P(X1)}}{\frac{P(X0)}{1-P(X0)}}$	Ile razy szanse zajścia są większe w grupie x1=1 niż x1=0
model logistyczny	$P(X) = \frac{1}{1+e^{-(\alpha+\beta X)}}$ $logit(A = a1) = log(odds(A = a1))$ $= log(P(A = a1)/(1 - P(A = a1)))$ $= \alpha + \beta X$	funkcja logitowa - do przekształcania prawdopodobieństwa na logarytm ilorazu szans:
model EVW	$logit\ P(X) = \alpha + \beta E + \gamma V + \delta W$ (wektorowy zapis)	E- efekt główny (wystawienie na ryzyko) V - potencjalnie zakłócające, W - potencjalnie modyfikujące efekty główne.
model EVW dla par połączonych	$logit\ P(X) = \alpha + \beta E + \gamma 1 V 1 +$ $+ \gamma 2 V 2 + \delta W$	Dodane zmienne 0-1 oznaczające status parowania przez niektóre zmienne C (control) np. płeć oznaczone V1

test zgodności Hosmera Lemeshowa	$$H = \sum_{i=1}^{G} \frac{(O(1i)-E(1i))^2}{E(1i)} + \frac{(O(0i)-E(0i))^2}{E(0i)}$$ ma rozkład X2 z G-2 stopniami swobody, gdzie G to liczba grup (najczęściej 10)	Test sprawdza, czy przewidywane liczby zdarzeń w podgrupach zbioru danych zgadzają się z rzeczywiście obserwowaną liczbą zdarzeń. Umożliwia stwierdzenie czy regresja logistyczna zwraca prawdopodobieństwo
krzywa ROC	oś x -> 1- specificity= 1-TN/(TN+FP) oś y -> sensivitity = (TP)/(TP+FN)	umożliwia jednoczesną ocenę czułości i swoistości modelu. Im krzywa ROC jest bardziej wypukła tym model jest lepszy (większe prawd op., że dobrze ustawimy próg odcięcia)

procedury Logistic

PROC LOGISTIC Statement Options	
Option	Description
Input/Output Data Set Options	
COVOUT	= [illegible]
DATA=	zbiór wejściowy
INEST=	
INMODEL=	
NOCOV	
OUTDESIGN=	
OUTDESIGNONLY	
OUTEST=	
OUTMODEL=	
Response and CLASS Variable Options	
DESCENDING	[illegible]
NAMELEN=	
ORDER=	
TRUNCATE	
Displayed Output Options	
ALPHA=	
NOPRINT	
PLOTS	(ONLY)=all → [illegible]
SIMPLE	
Large Data Set Option	
MULTIPASS	
Control of Other Statement Options	
EXACTONLY	
EXACTOPTIONS	
ROCOPTIONS	dodatkowe opcje do wykresu krzywej ROC


2. Zapisz relacje pomiędzy Odds & prawdopodobieństwem

Odds (zdarzenie) =

Prawdopodobieństwo (zdarzenie) =

3.Jeśli prawdopodobieństwo wystąpienia zdarzenia

4.Wyniki estymacji funkcji logitowej są następujące

dychotomicznej X. Ile wynoszą Odds dla

5 . Wyjaśnij różnicę w metodach estymacji

modeli regresji

2. Zapisz relację między odds a prawdopodobieństwem

p=odds/(odds+1)

odds=p/(1-p)

3. Jeżeli prawdopodobieństwo wystąpienia zdarzenia c = 0,6 jakie jest odds

odds=0,6/(1-0,6)=1,5

4. Wynik estymacji funkcji logistycznej są następujące: p(x,y)=1,9386 * y + 0,231* x dla dychotomicznej zmiennej x. Ile wynosi odds dla y=1|x=1 względem y=1|x=0?

$$OR = \frac{Odds\ dla\ X1}{Odds\ dla\ X0} = \frac{\frac{P(X1)}{1-P(X1)}}{\frac{P(X0)}{1-P(X0)}}$$

p1 = 1/(1+e^(-[1,9386 * 1 + 0,231* 1])) = 0,89749

p0 = 1/(1+e^(-[1,9386 * 1 + 0,231* 0])) = 0,87420

Odds1 = p1/(1-p1) = 8,754781

Odds0 = p0/(1-p0) = 6,949016

OR = Odds1/Odds0 = 1,259859 -

OR = TO SAMO WYJDZIE JAK OBLICZYMY EXP(0.231)

czyli exp(z1-z2), gdzie z to logit tego modelu

5. Wyjaśnij różnicę w metodach estymacji MLM i RMLM stosowanych do estymacji modeli regresji

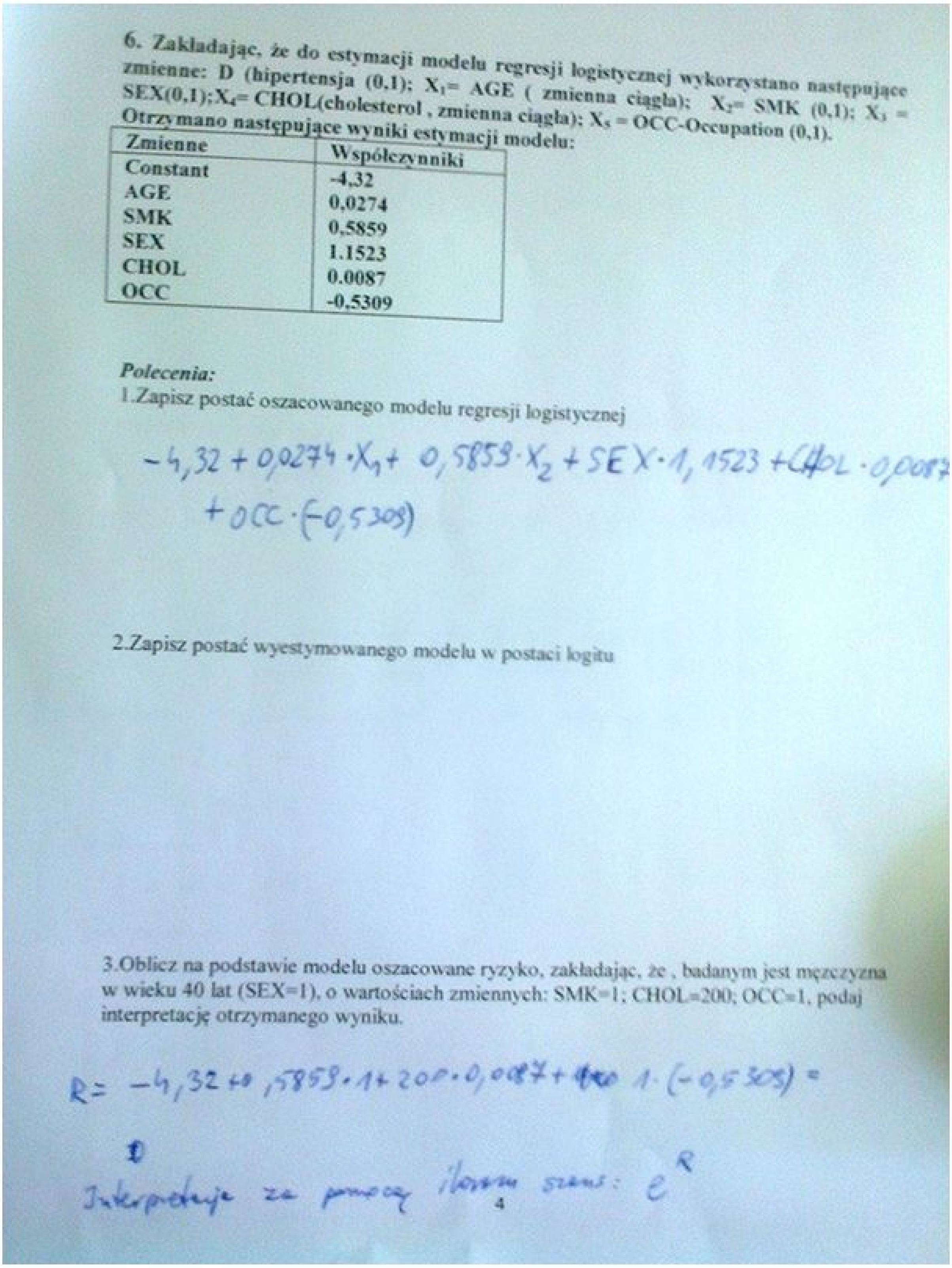
MLM - https://en.wikipedia.org/wiki/Multilevel_model ?

Maximum Likelihood Method - a method of estimating the parameters of a statistical model given observations, by finding the parameter values that maximize the likelihood of making the observations given the parameters.

Suppose there is a sample $x_1, x_2, \ldots, x_n$ of n independent and identically distributed observations, coming from a distribution with an unknown probability density function $f_0(\cdot)$. It is however surmised that the function f_0 belongs to a certain family of distributions $\{ f(\cdot|\ \theta), \theta \in \Theta \}$ (where θ is a vector of parameters for this family), called the parametric model, so that $f_0 = f(\cdot|\ \theta_0)$. The value θ_0 is unknown and is referred to as the *true value* of the parameter vector. It is desirable to find an estimator theta which would be as close to the true value θ_0 as possible. Either or both the observed variables x_i and the parameter θ can be vectors.

Robust Maximum Likelihood Method - gdzie założeniem również jest że dane mają wielowymiarowy rozkład normalny ale kurtoza jest mniejsza lub większa niż 3, [normalna dla normalnego =3]

-



6. Zakładając, że do estymacji modelu regresji logistycznej wykorzystano następujące zmienne: D (hipertensja (0,1); X_1= AGE (zmienna ciągła); X_2= SMK (0,1); X_3 = SEX(0,1);X_4= CHOL(cholesterol , zmienna ciągła); X_5 = OCC-Occupation (0,1). Otrzymano następujące wyniki estymacji modelu:

Zmienne	Współczynniki
Constant	-4,32
AGE	0,0274
SMK	0,5859
SEX	1.1523
CHOL	0.0087
OCC	-0,5309

Polecenia:

1.Zapisz postać oszacowanego modelu regresji logistycznej

$-4,32 + 0,0274 \cdot X_1 + 0,5859 \cdot X_2 + SEX \cdot 1,1523 + CHOL \cdot 0,0087$
$+ OCC \cdot (-0,5309)$

2.Zapisz postać wyestymowanego modelu w postaci logitu

3.Oblicz na podstawie modelu oszacowane ryzyko, zakładając, że , badanym jest mężczyzna w wieku 40 lat (SEX=1), o wartościach zmiennych: SMK=1; CHOL=200; OCC=1, podaj interpretację otrzymanego wyniku.

$R = -4,32 + 0,5859 \cdot 1 + 200 \cdot 0,0087 + 1 \cdot (-0,5309) =$

Interpretacja za pomocą ilorazu szans: e^R

4

1. **zapisz postać oszacowanego modelu regresji logistycznej**

pi= 1/(1+e^(-[-4,32 + 0,0274 * AGE + 0,5859 * SMK + 1.1523 * SEX + 0,0087 * CHOL - 0,5309 * OCC]))

2. **zapisz postać wyestymowanego modelu w postaci logitu**

zi= -4,32 + 0,0274 * AGE + 0,5859 * SMK + 1.1523 * SEX + 0,0087 * CHOL - 0,5309 * OCC

3. **oblicz na podstawie modelu oszacowane ryzyko, zakładając że badanym jest mężczyzna w wieku 40 lat o wartości zmiennych SMK=1 CHOL=200 OCC=1. podaj interpretację**

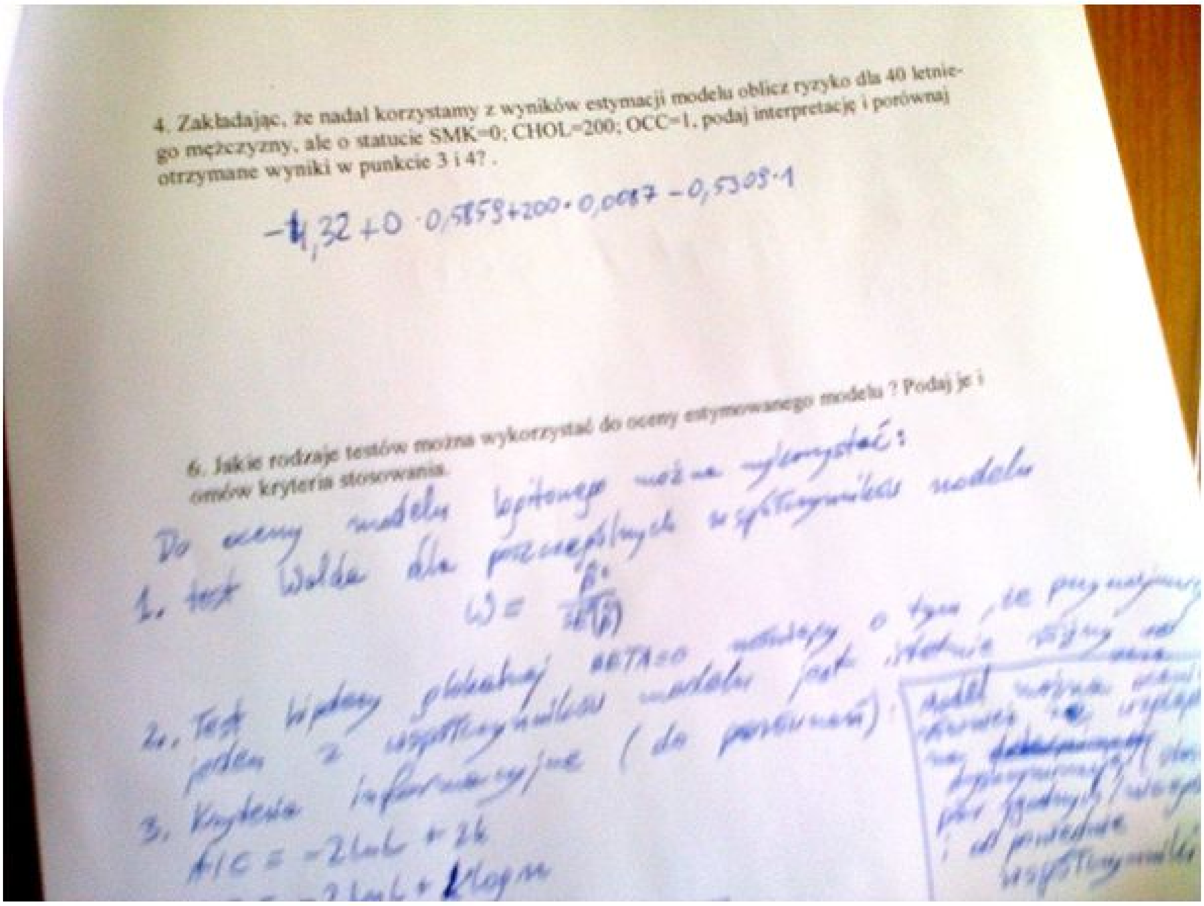
R1 = -4,32 + 0,0274 * 40 + 0,5859 * 1 + 1,1523 * 1 + 0,0087 * 200 - 0,5309 * 1 = -0,2767

Ryzyko można interpretować jedynie w odniesieniu do innego ryzyka. Można mówić że ryzyko dla jednostki jest większe niż dla innej jednak samo ryzyko nie jest prawdopodobieństwem ?

A to nie będzie 1/(1+e^(-[R1])) = 0,431263

Interpretacja: Ryzyko zachorowania dla wybranej osoby wynosi 43%

(też mi się tak wydaje)



4. Zakładając, że nadal korzystamy z wyników estymacji modelu oblicz ryzyko dla 40 letniego mężczyzny, ale o statucie SMK=0; CHOL=200; OCC=1, podaj interpretację i porównaj otrzymane wyniki w punkcie 3 i 4?.

$-4,32 + 0 \cdot 0,5859 + 200 \cdot 0,0087 - 0,5309 \cdot 1$

6. Jakie rodzaje testów można wykorzystać do oceny estymowanego modelu? Podaj je i omów kryteria stosowania.

Do oceny modelu logitowego można wykorzystać:

1. test Walda dla poszczególnych współczynników modelu $W = \frac{\beta_i}{SE(\beta_i)}$

2. Test wspólny globalnej ... współczynników modelu (do porównań)

3. Kryteria informacyjne

$AIC = -2\ln L + 2k$

$-2\ln L + k\log n$

4. zakładając że nadal korzystamy z wyników estymacji modelu oblicz ryzyko dla 40 letniego mężczyzny w wieku 40 lat o wartości zmiennych SMK=0

CHOL=200 OCC=1. podaj interpretację i porównaj otrzymane wyniki w punkcie 3 i 4
R2 = -4,32 + 0,0274 * 40 + 0,5859 * 0 + 1,1523 * 1 + 0,0087 * 200 - 0,5309 * 1 = -0,8626

RR=R2/R1=-0,8626/-0,2767=3,1175

Oznacza to że Ryzyko dla drugiej osoby jest ponad 3 krotnie mniejsze niż dla osoby pierwszej ?

A to nie będzie 1/(1+e^(-[R2])) = 0,297
Interpretacja: Ryzyko zachorowania dla wybranej osoby wynosi 30%
RR= P1/P2= 1,45
Prawdopodobieństwo zdarzenia dla 1 osoby jest o 45% wyższe niz dla 2 osoby
(popieram)

6. Jakie rodzaje testów można wykorzystać do oceny estymowanego modelu? Podaj je i omów kryteria stosowania
test Walda dla poszczególnych współczynników modelu
Kryteria informacyjne: (im mniejsze tym lepiej)

- funkcja logarytmu z funkcji wiarygodności -2lnL
- AIC = -2lnL + 2k (doklada kare za liczbe zmiennych w modelu - k)
- BIC = -2lnL + k * ln(n) (doklada kare za liczbe zmiennych w modelu - k i liczbe obserwacji (n)

AUC - Area Under Curve - Pole pod krzywą ROC
Testy:

- test stosunku wiarygodności
- test punktowy
- test Walda

Jeżeli wielkość sprawdzianu z tych testów jest statystycznie istotna, daje to podstawę do przyjęcia?? hipotezy zerowej, że przynajmniej jeden ze współczynników w populacji generalnej jest istotnie różny od zera

Kartkówka 2

……………albumu……………………grupa

Zadanie 1 (5 punktów) MNW w estymacji modelu regresji logistycznej

What is the probability of our data—seeing 7 heads in 10 coin tosses—as a function p? The number of heads in 10 coin tosses is a binomial random variable with N=10 and p=(unknown) p.

$$\therefore P(7\text{ heads}) = \binom{10}{7} p^7 (1-p)^3 = \frac{10!}{7! * 3!} p^7 (1-p)^3$$

Metodą największej wiarygodności wyznacz wartość parametru p.
Ile wynosi: -2 (loglikelihood) ?

<u>Rozwiazanie :</u>

1. Take the log of the likelihood function.

$$Likelihood = \binom{10}{7} p^7 (1-p)^3 = \frac{10!}{7! * 3!} p^7 (1-p)^3$$

$$\log Likelihood = \log \frac{10!}{7! * 3!} + 7 \log p + 3 \log(1-p)$$

2. Take the derivative with respect to p.

$$\frac{d}{dp} \log Likelihood = 0 + \frac{7}{p} + \frac{-3}{1-p}$$

Jog your memory→

*derivative of a constant is 0

*derivative 7f(x)=7f '(x)

*derivative of log x is 1/x

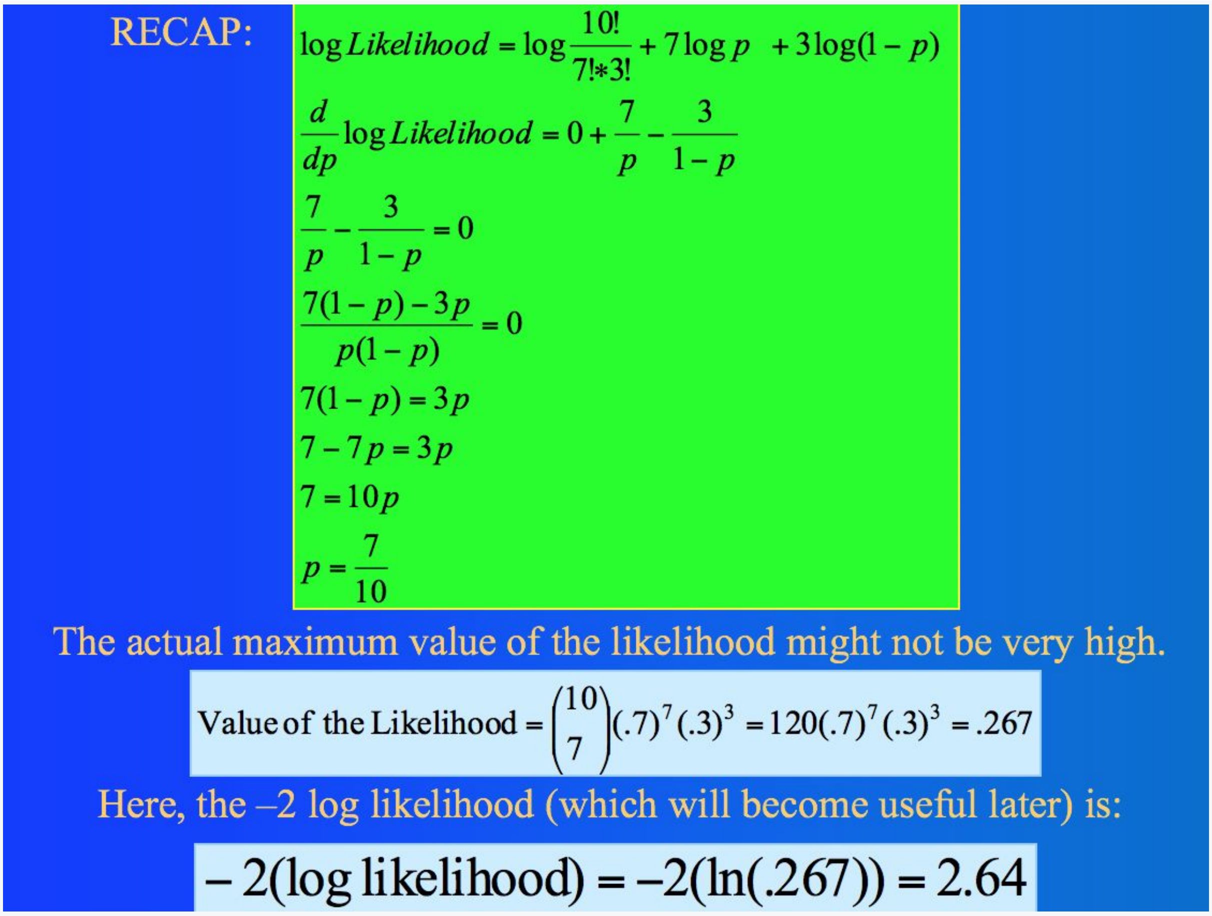
*chain rule

3. Set the derivative equal to 0 and solve for p.

$$\frac{7}{p} - \frac{3}{1-p} = 0 \longrightarrow \frac{7(1-p) - 3p}{p(1-p)} = 0 \longrightarrow 7(1-p) = 3p$$

$$7 - 7p = 3p \longrightarrow 7 = 10p$$

$$p = \frac{7}{10}$$



Max(log(p^7*(1-p)^3)))
->
d/dp(log(p^7 (1 - p)^3)) = (7 - 10 p)/(p - p^2)
->
miejsce zerowe P=0,7
->
Max(log(p^7*(1-p)^3))) dla p=0,7

1 wyznacz z definicji wartość odds ratio i podaj jego interpretację teoretyczną
https://www.google.pl/url?sa=t&source=web&rct=j&url=http%3A%2F%2Fweb.stanford.edu%2F~kcobb%2Fhrp261%2Flecture4.ppt&ved=0ahUKEwj3gfGOzNTLAhVlMnIKHXotDQoQFggMAE&usg=AFQjCNFtVPlqLW8SITc8LZW_ooQGOX9Nbw&sig2=gU3hM7x7TbJGNTXakoqzbw
slajd 28 i 29 - dokładnie przekopiować tę tabelę

	Exposure=1	Exposure=0
Disease = 1	$P(D/E)=\frac{e^{\alpha+\beta_1}}{1+e^{\alpha+\beta_1}}$	$P(D/\sim E)=\frac{e^{\alpha}}{1+e^{\alpha}}$
Disease = 0	$P(\sim D/E)=\frac{1}{1+e^{\alpha+\beta_1}}$	$P(\sim D/\sim E)=\frac{1}{1+e^{\alpha}}$

Odds Ratio for simple 2x2 Table

$$OR=\frac{\frac{e^{\alpha+\beta_1}}{\cancel{1+e^{\alpha+\beta_1}}}\times\frac{1}{\cancel{1+e^{\alpha}}}}{\frac{1}{\cancel{1+e^{\alpha+\beta_1}}}\times\frac{e^{\alpha_1}}{\cancel{1+e^{\alpha}}}}=\frac{e^{\alpha+\beta_1}}{e^{\alpha}}=e^{(\alpha+\beta_1)-\alpha}=e^{\beta_1}$$

(courtesy *Hosmer and Lemeshow*)

To samo inaczej

NiOk

1 wyznacz z definicji wartość odds ratio i podaj jego interpretację teoretyczną

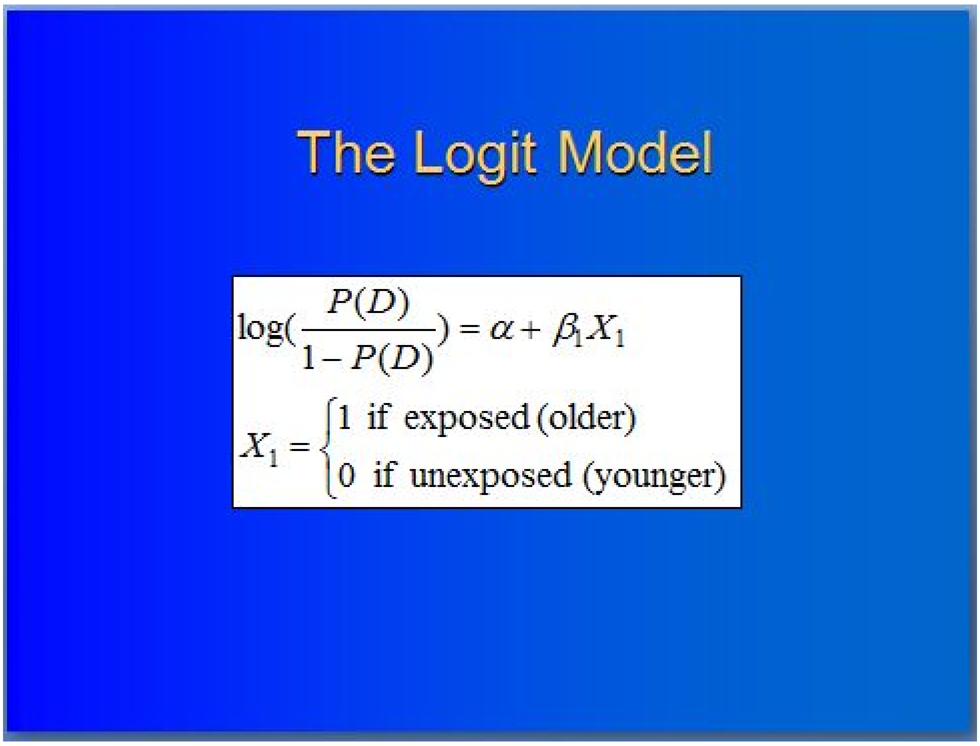
$$1-P(X)=1-\frac{1}{1+e^{-(\alpha+\sum\beta_i X_i)}}=\frac{e^{-(\alpha+\sum\beta_i X_i)}}{1+e^{-(\alpha+\sum\beta_i X_i)}}$$

$$\frac{P(X)}{1-P(X)}=\frac{\dfrac{1}{1+e^{-(\alpha+\sum\beta_i X_i)}}}{\dfrac{e^{-(\alpha+\sum\beta_i X_i)}}{1+e^{-(\alpha+\sum\beta_i X_i)}}}=e^{(\alpha+\sum\beta_i X_i)}$$

Interpretacja teoretyczna:
Oznacza to że szansa dla osób które mają parametr B1=1 jest e^B1 większa niż dla osób które ten parametr mają na poziomie B1=0

?

2. Zapisz postać modelu w postaci logitu



log(P/[1-P])=a+**BX**

Przykład:

Grupa odniesienia= ***Adenocarcinoma***

Dwa porównania:

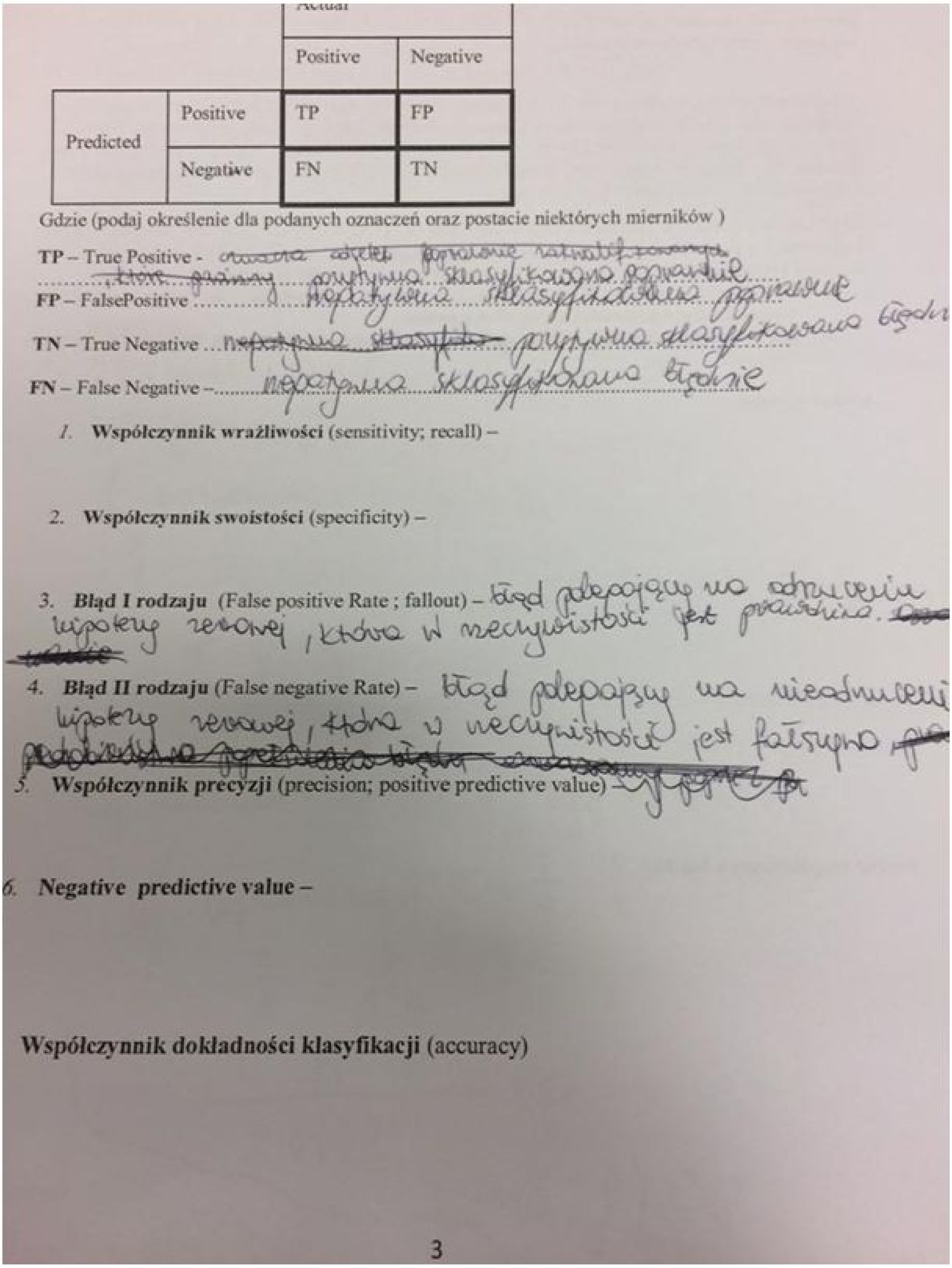
1. *Adenosquamous (D=1) vs Adenocarcinoma (D=0)*
2. *Inne (D=2) vs Adenocarcinoma (D=0)*

Korzystając z danych tablicy mamy:

CRUDE ODDS RATIO

$$\hat{OR}_{1 vs 0} = \frac{77 \times 34}{109 \times 11} = 2.18$$

$$\hat{OR}_{2 vs 0} = \frac{77 \times 39}{109 \times 18} = 1.53$$



		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Gdzie (podaj określenie dla podanych oznaczeń oraz postacie niektórych mierników)

TP – True Positive - ~~ilość obiektów poprawnie zaklasyfikowanych~~ które pozytywna sklasyfikowano poprawnie

FP – FalsePositive ... negatywna sklasyfikowano pozytywnie

TN – True Negative ... ~~negatywna sklasyf~~ pozytywna sklasyfikowano błędnie

FN – False Negative – negatywna sklasyfikowano błędnie

1. **Współczynnik wrażliwości** (sensitivity; recall) –
2. **Współczynnik swoistości** (specificity) –
3. **Błąd I rodzaju** (False positive Rate ; fallout) – błąd polegający na odrzuceniu hipotezy zerowej, która w rzeczywistości jest prawdziwa.
4. **Błąd II rodzaju** (False negative Rate) – błąd polegający na nieodrzuceniu hipotezy zerowej, która w rzeczywistości jest fałszywa
5. **Współczynnik precyzji** (precision; positive predictive value) –
6. **Negative predictive value** –

Współczynnik dokładności klasyfikacji (accuracy)

3

TP - true positive - liczba poprawnie przewidzianych sukcesów
FP - False positive - liczba niepowodzeń przewidywanych niepoprawnie jako sukces
TN - true Negative - liczba poprawnie przewidywanych niepowodzeń
FN - False negative - liczba sukcesów przewidywanych niepoprawnie jako niepowodzenie
Wspólczynnik wrażliw§ości (sensivty, recal) - TP/(TP+FN) <- miara wykrywania sukcesów, % dobrze przewidzianych sukcesów w ogóle sukcesów w danych

W**spółczynnik swoistości (specifity)** - TN/(TN+FP) <- miara wykrywania niepowodzeń, % niepowodzeń dobrze przewidziancyh jako negatywy (z ogółu negatywów w danych)
błąd 1 rodzaju - FP/(FP+TN) <- odsetek fałszywych alarmów,
A nie będzie to: błąd polegający na odrzuceniu hipotezy zerowej, która w rzeczywistości nie jest fałszywa **-> to to samo**
błąd 2 rodzaju - FN/(FN+TP)
A nie będzie to: błąd polegający na nieodrzuceniu hipotezy zerowej, która jest w rzeczywistości fałszywa. **-> to to samo**
wspólczynnik precyzji (precision ; positive predicted value) TP/(TP+FP)
http://www.uta.fi/sis/tie/tl/index/Rates.pdf
negative predictive value - TN/(TN+FN)
współczynnik dokładności klasyfikacji (TP+TN)/(P+N)

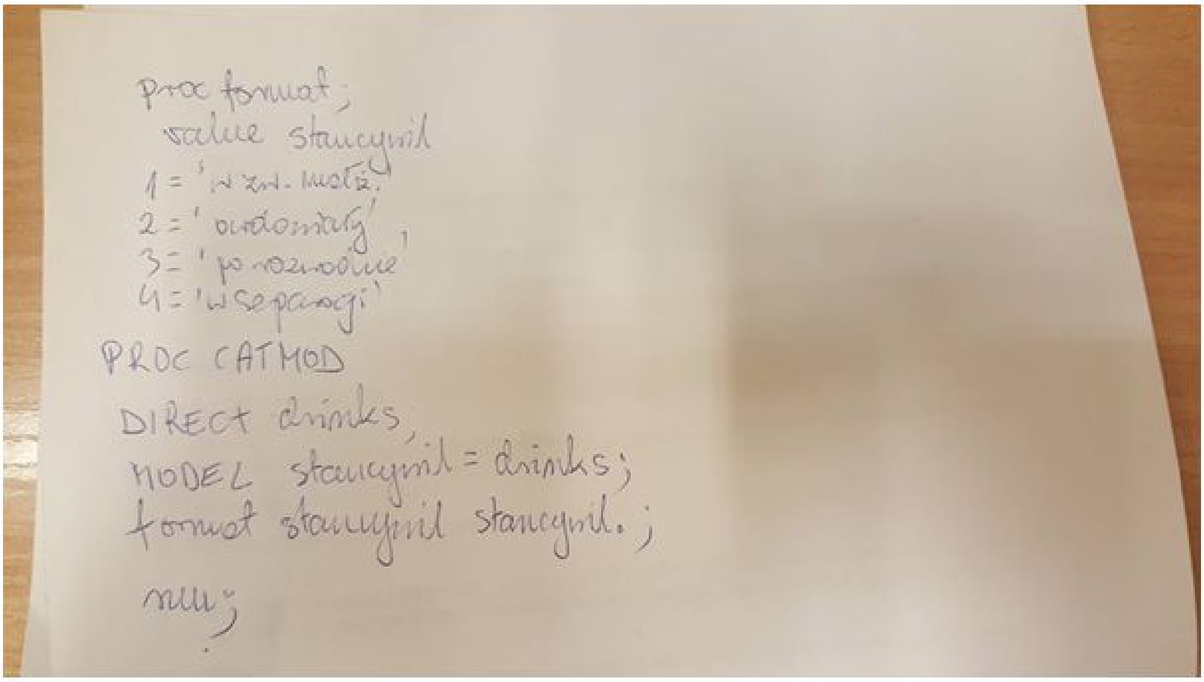
Zadanie 5 (4 punkty) Jeśli założenie proporcjonalnych szans jest nie do utrzymania (iloraz szans zmienia się w zależności od kategorii), możemy dopasować zbiór równań regresji logistycznej, które będą posiadać zarówno indywidualne współczynniki kierunkowe jak i wyrazy stałe. Ten alternatywny model odzwierciedla tzn. odpowiedzi Y są w swojej naturze nominalne. W takim wypadku załóżmy, że mamy k nieuporządkowanych kategorii odpowiedzi wraz ze związanymi z nimi *prawdopodobieństwami kategorii* $p_j = P(Y=j|X)$, $j = 1,\dots,k$. Wtedy zbiór $k-1$ *uogólnionych lub wielomianowych logitów* ma postać

$\log\left(\frac{p_j}{p_k}\right) = \beta_{0j} + \beta_{1j}X, j = 1, \dots k-1$ gdzie każda kolejna kategoria od 1 do k-1 jest porównywana z kategorią k lub ostatnią. Jak widać, system ten pozwala na istnienie $k-1$ różnych współczynników kierunkowych i wyrazów stałych. Załóżmy losową próbę n=500 dorosłych, którzy zostali spytani o ich stan cywilny (Y) i liczbę napojów alkoholowych spożywanych w ciągu tygodnia (X). Kategorie odpowiedzi nominalnych dla Y są następujące: 1, w związku małżeńskim; 2. Owdowiały ; 3, po rozwodzie; 4, w separacji .

Aby określić zależność pomiędzy odpowiedziami Y i X, z wykorzystaniem modelu regresji *logistycznej o postaci wielomianu* , gdzie mamy dwie zmienne: *stancywil* (Y) – stan cywilny – to *zmienna zależna, a drinks (X)* - ilość napoi alkoholowych – to zmienna niezależna.

1. **Napisz kod programu SAS na estymację modelu , zakładając , że informacje o 500 respondentów wprowadza się do SAS z ankiet .**
2. **Jak w postaci równania należy zapisać postać trzech uogólnionych logitów (w postaci równań):**

Kod programu



proc format;
value stancywil
1 = 'w zw. małż.'
2 = 'owdowiały'
3 = 'po rozwodzie'
4 = 'w separacji'
PROC CATMOD
DIRECT drinks,
MODEL stancywil = drinks;
format stancywil stancywil.;
run;

1. **Kod w SAS**

```
proc format;
value stancywil
1=”w związku”
2=”owdowiały”
3=”po rozwodzie”
4= “w separacji”;

proc catmod;
	direct drinks;
	model stancywilny = drinks;
	format stancywilny stancywil.;
run;
```

2. jakw postaci wielomianu należy zapisać postać trzech uogólnionych logitów (w postaci równań)

$$\log\left(\frac{p_1}{p_4}\right) = \beta_{01} + \beta_{11}X$$

$$\log\left(\frac{p_2}{p_4}\right) = \beta_{02} + \beta_{12}X$$

$$\log\left(\frac{p_3}{p_4}\right) = \beta_{03} + \beta_{13}X$$

$$\frac{p_1}{p_4} + \frac{p_2}{p_4} + \frac{p_3}{p_4} = e^{\beta_{01}+\beta_{11}X} + e^{\beta_{02}+\beta_{12}X} + e^{\beta_{03}+\beta_{13}X}$$

$$p_1 + p_2 + p_3 = p_4(\ldots)$$

$$p_1 + p_2 + p_3 + p_4 = 1$$

$$p_1 + p_2 + p_3 = 1 - p_4$$

$$1 - p_4 = p_4[.]$$

$$1 = p_4[.] + p_4$$

$$1 = p_4([.] + 1)$$

$$p_4 = \frac{1}{e^{\beta_{01}+\beta_{11}X} + e^{\beta_{02}+\beta_{12}X} + e^{\beta_{03}+\beta_{13}X} + 1}$$

czy to może chodzić o coś takiego?

czy bardziej o:

$$1: \log\left(\frac{P(X=1)}{P(X=2 \vee X=3 \vee X=4)}\right) = \log\left(\frac{P(X=1)}{P(X=2)+P(X=3)+P(X=3)}\right)$$

$$2: \log\left(\frac{P(X=2)}{P(X=1 \vee X=3 \vee X=4)}\right) = \log\left(\frac{P(X=2)}{P(X=1)+P(X=3)+P(X=4)}\right)$$

......