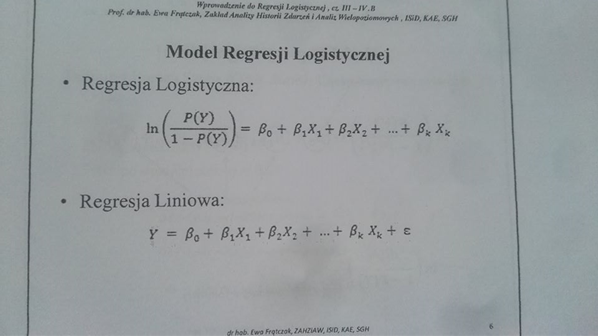
**1. Porównać założenia regresji logistycznej i liniowej. Które założenia liniowe nie są spełnione przez logistyczną.**

**Model regresji logistycznej założenia:**

* estymowany jest model proporcjonalnych ODDS

Regresja liniowa- zmienna wynikowa **CIĄGŁA (skala ilościowa)**

Regresja logistyczna - zmienna wynikowa **DYCHOTOMICZNA przyjmująca wartości (0,1)**

****

**Nie należy używać regresji liniowej w przypadku dychotomicznych zmiennych wynikowych:**

* gdy modelujemy Y bezpośrednio i Y jest zmienną dychotomiczną zostaną naruszone założenia regresji liniowej
* przewidziane wartości Y mogą przyjmować wartości spoza (0,1)

**Wspólne cechy regresji liniowej i logistycznej:**

* należą do jednej rodziny uogólnionych modeli liniowych GLM
* operują skalą logarytmiczną co pozwala na zastosowanie modelu liniowego

**W regresji logistycznej:**

1. Yi pochodzą z rozkładu Bernoulliego lub dwumianowego (ni, µi)

2. Obserwacje (Yi) są niezależne

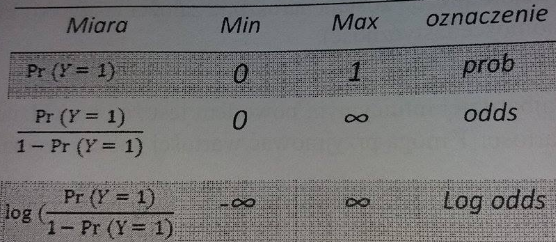
3. Log odds P(Yi=1) lub logit P(Yi=1) jest liniową funkcją zmiennych

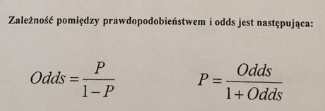
Własności estymatorów obu metod.







4. Zależność między prawdopodobieństwem, ilorazem szans oraz logarytmem ilorazu szans



**2. Obliczyć na podstawie tablic kontyngencji:**  
- ryzyko (RR)

RR= 1(X)/ 0(X)

(X)= 1/1+e^-[( α+β1(CAT)+ β2(WIEK)+ β3(ECG)]

Gdzie CAT, WIEK, ECG to wartości naszych zmiennych

- ryzyko względne,

**RW=P2/p1 lub odds dla 1/odds dla 0**

-ODDS RATIO

**OR = ad/ bc**

Proszę (policzone na wymyślonych danych):

Prawdopodobieństwo, że Jasiu pójdzie jutro do szkoły: 0,3 (p1)

Prawdopodobieństwo, że Małgosia pójdzie jutro do szkoły: 0,7 (p2)

Szanse, że jasiu pójdzie jutro do szkoły: p1 / 1-p1 = 0,3 / 0,7 = 3:7 (O1)

Szanse, że małgosia pójdzie jutro do szkoły: p2 / 1 - p2 = 0,7 / 0,3 = 7:3 (O2)

Iloraz szans Jasia i Małgosi:

O1 / O2 = (3/7) / (7/3) = 0,18

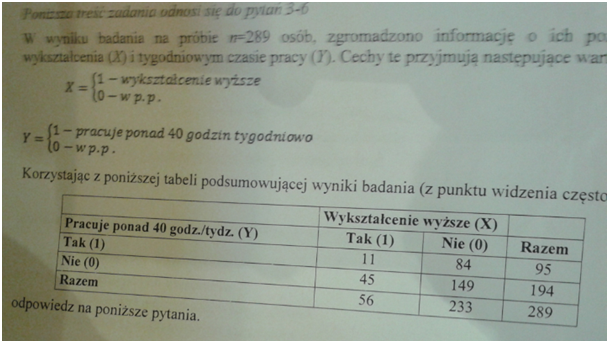
Szanse, że jasiu pójdzie jutro do szkoły stanowią 0,18 szans, że Małgosia pójdzie jutro do szkoły

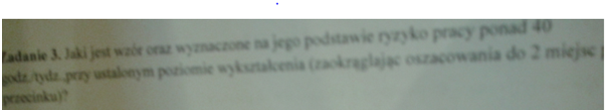
W drugą stronę (iloraz szans Małgosi i Jasia):

O2 / O1 = (7/3)(3/7) = 5,44

Szanse, że Małgosia pójdzie jutro do szkoły wynoszą 5,44 szans jasia (są o 444% większe!)

Zadanie z egzaminu:





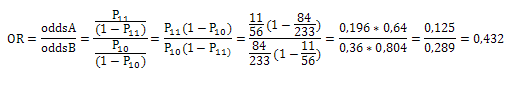
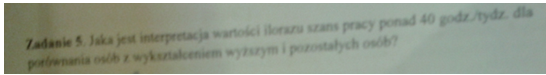
= 0,2 (zaokrąglone) - wykszt wyższe

- przeciwny przypadek



RR=P(a)/P(b)=0,2/0,36 = 0,56

Osoby z wykształceniem wyższym mają o 44% mniejsze ryzyko na prace powyżej 40h w tygodniu niz….



Osoby z wyksztalceniem wyz. maja sznsa na prace powyżej 40h tyg o 57% niższe niż osoby bez wykształcenia wyższego

**3. Wydruk z oszacowanymi parametrami i przedziałami ufności.(Oceny parametrów i )**  
Obliczyć:  
- punktową ocenę ODDS RATIO - exp(beta)  
- przedziały ufności - exp(niższy przedział) exp(wyższy przedział)

Gdyby była tablica Analysis of Maximum Likelihood Estimates to:

-punktowa ocena ODDS RATIO: exp(beta)

- przedziały ufności dla współczynnika regresji: Estimate +/- 1,96 \* Standard Error

- przedziały ufności dla ilorazu szans: exp(niższy przedział) exp(wyższy przedział)

**4. Kody do napisania na kartce: MA KTOŚ OGÓLNE POSTACI KODÓW?**  
**- binarna,**

**(Ogólna postać kodów - bez dodatkowych opcji, dla przykladowej zmiennej celu Happy\_bin i objaśnijących GNDR i HINCFEL- może ktoś potwierdzić?)**

|  |
| --- |
| **Binarna:**  **PROCLOGISTIC DATA= DANE**  **CLASS GNDR HINCFEL ;**  **MODELHappy\_bin (Event = '1')= GNDR HINCFEL /**  **LINK=LOGIT;**  **RUN;**    **Wielomianowa:**  **PROCLOGISTICDATA= DANE**  **CLASS GNDR HINCFEL ;**  **MODEL Happy\_uporz = GNDR HINCFEL /**  **LINK=GLOGIT;**  **RUN;**    **Uporządkowana:**  **PROCLOGISTICDATA= DANE**  **CLASS GNDR HINCFEL ;**  **MODEL RESPONSE (DESCEDNING)= GNDR HINCFEL /**  **LINK=LOGIT;**  **RUN;** |

**dodatkowe opcje (do każdej):**

**CLASS GNDR (param=ref ref=’2’) HINCFEL (param= effect) ; - kodowanie**

**MODELHappy\_bin (Event = '1')= GNDR HINCFEL /**

**SELECTION = none/stepwise/forward/backward - metoda selekcji**

**scale = none -statystyki dewiancji**

**aggregate - statystyki Pearsona**

**lackfit - test Hosmera-Lemeshowa**

**expb - oceny dopasowanych ilorazów szans**

**ctable pprob=.5 -tablica klasyfikacji dla każdego progu odcięcia**

rsquare - r kwadrat i skorygowane r kwadrat

clparm=both - wyświetlenie przedziałów ufności dla parametrów

clodds=both - wyświetlenie przedziałów ufności dla ilorazów szans

alpha=**0.05**; - poziom istotności

reschi= \_reschi1 resdev= \_resdev1 - reszty dewiancji i Pearson chi-square

predicted= **\_predicted1**; - zmienna z wartościami teoretycznymi

clparm=WALD - przedziały ufności Walda dla parametrów

clodds=WALD - przedziały ufności Walda dla ilorazów szans

UNITS wiek=5 - pogrupowane co ileś jednostek

**ROC:**

**ods graphics on;**

**proc logistic data=hsb2 plots(only)=roc(id=obs);**

**class prog (ref='1') /param = ref;**

**model hiwrite(event='1') = female prog read math**

**/\*Model regresji binarnej\*/**

PROC LOGISTIC DATA=WORK.SORTTempTableSorted

PLOTS(ONLY)=ODDSRATIO

PLOTS(ONLY)=ROC;

CLASS PPLFAIR (PARAM=REF) TRSTPLT (PARAM=REF) HEALTH

(PARAM=REF) GNDR (PARAM=REF) AGEA (PARAM=REF) HINCFEL

(PARAM=REF) geduyrs (PARAM=REF) gmaritalb (PARAM=REF);

MODEL rstflife (Event = '1')=PPLFAIR TRSTPLT HEALTH GNDR AGEA HINCFEL

geduyrs gmaritalb /

SELECTION=STEPWISE

SLE=0.05

SLS=0.05

INCLUDE=0

INFLUENCE

LACKFIT

AGGREGATE SCALE=NONE

RSQUARE

LINK=LOGIT

CLPARM=BOTH

CLODDS=BOTH

ALPHA=0.05

;

OUTPUT OUT=WORK.PREDLOGREGPREDICTIONS\_0001(LABEL="Statystyki i

prognozy regresji logistycznej dla WORK.PREDLOGREGPREDICTIONS\_0000")

PREDPROBS=INDIVIDUAL

UPPER=upper\_rstflife

LOWER=lower\_rstflife ;

RUN;

QUIT;

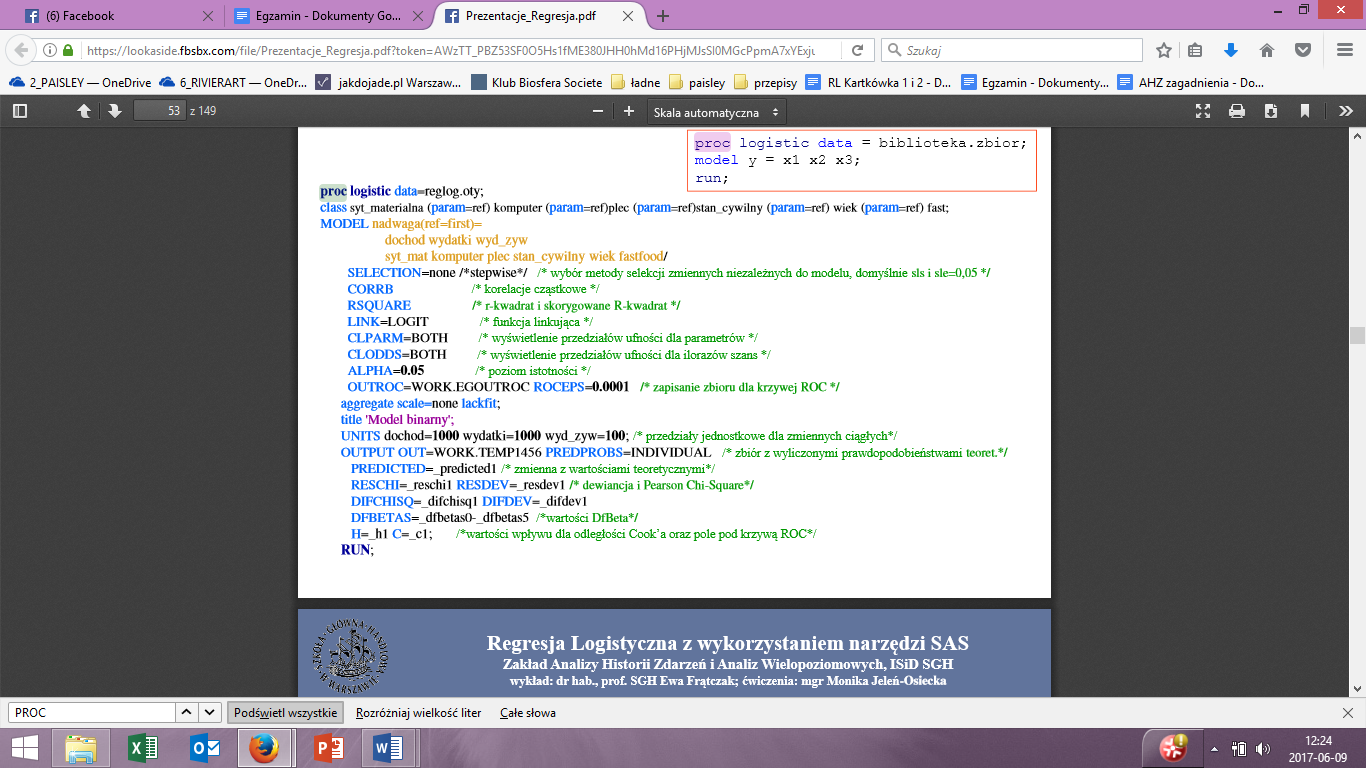
TITLE;

TITLE1 "Prognozy analizy regresji";

PROC PRINT NOOBS DATA=WORK.PREDLOGREGPREDICTIONS\_0001

;

RUN;



**- wielomianowa**

**proc** **logistic** nosimple data=reg\_wiel.syt\_mater2;

class samochod (param=ref) wyksztalcenie (param=ref) stan\_cyw (param=ref)

TYP (param=ref) plec (param=ref) syt\_materialna (param=ref);

model grupa (ref='0')=doch\_rozp wyd\_total samochod wyksztalcenie stan\_cyw

TYP plec syt\_materialna/

selection=none

expb scale=none aggregate

corrb

rsquare

link=glogit

clparm=both

clodds=both

alpha=**0.05**;

output out=reg\_wiel.wyniki p=predict l=lower u=upper xbeta=logit stdxbeta=selogit /

alpha=**0.05**;

**run**;

**- uporządkowana**

**proc** **logistic** nosimple data=propodds.syt\_mater2 descending;

class KLM (param=ref) l\_ludnosci (param=ref) TYP (param=ref)

l\_pokoi (param=ref) komputer (param=ref)

telefon (param=ref) samochod (param=ref) plec (param=ref)

stan\_cyw (param=ref) wyksztalcenie (param=ref);

model syt\_materialna = KLM l\_ludnosci TYP l\_pokoi komputer telefon

samochod plec stan\_cyw wyksztalcenie

DOCHG WYDG WIEK DG905 DG906 W01 W02 W03 W04 W05

W06 W07 W08 W09 W10 W11 W12 W13 /

lackfit expb scale=none aggregate selection=stepwise

RSQUARE

LINK=LOGIT

CLPARM=BOTH

CLODDS=BOTH

ALPHA=**0.05**;

output out=propodds.wyniki p=predict l=lower u=upper xbeta=logit stdxbeta=selogit / alpha=**.05**;

**run**;

* zgrupowanie wieku w jednostce po 10 lat,

UNITS wiek=10 - pogrupowane co ileś jednostek

* kod na test homera lebieszowa

**lackfit - test Hosmera-Lemeshowa**

Test zgodności Hosmera i Lemeshowa służy do interpretacji dopasowania

modelu do danych. Hipoteza zerowa H0 mówi o dobrym dopasowaniu modelu do danych.

Test zgodności Hosmera i Lemeshowa bazuje na grupowaniu obserwacji ze względu na wartości wyestymowanych prawdopodobieństw. Najczęściej wyróżnia się dwie strategie grupowania obserwacji.

1. Podział wszystkich obserwacji na g=10 grup, z czego pierwsza zawiera n1’=n/10 obserwacji charakteryzujących się najniższymi wartościami wyestymowanego prawdopodobieństwa, a ostatnia n10’=n/10 najwyższymi prawdopodobieństwami.
2. Drugi podział polega na uporządkowaniu obserwacji po wcześniejszym ustaleniu punktów granicznych zdefiniowanych jako k/10, gdzie k=1,2,3...9. Do danej grupy wpadają obserwacje, które odznaczają się wartością wyestymowanego prawdopodobieństwa zawierającą się między ustalonymi punktami granicznymi.

proc logistic data=my.mroz;

model inlf(desc) = kidslt6 age educ huswage city exper / lackfit;

run;

The LACKFIT option is what requests the Hosmer-Lemeshow (HL) test.

DO POTWIERDZENIA! <- tak, potwierdzam, też

**5. Był jakiś wydruk/kod i trzeba było się zastanowić czy coś trzeba zmienić** (na Walda i Units)

to z units to po prostu dopisać linijkę:  
UNITS dochod=1000;

a co do Walda, to mogą być liniki kodu

CLPARM=WALD - PRZEDZIAŁY UFNOŚCI PARAMETRÓW STATYSTYKĄ WALDA

CLODDS=WALD - WARUNKOWY ILORAZ SZANS STATYSTYKĄ WALDA

ALPHA=0.05

**6. Interpretacja testów na spełnienie założenia.**  
7. Do interpretacji:  
- miary dyskryminacji   
Celem dyskryminacji modelu, czyli badania w jakim stopniu model dobrze rozróżnia obiekty w obu grupach, jest wyliczenia procenta par zgodnych i niezgodnych. Innymi słowy celem dyskryminacji jest przypisanie jednostkom na podstawie zbudowanego modelu odpowiedniej kategorii zmiennej objaśnianej (w przypadku binarnej zmiennej celu model ma za zadanie określenie, czy dana jednostka zanotowała lub nie zajście badanego zdarzenia) i porównanie tak nadanych wartości teoretycznych z wartościami faktycznie zaobserwowanymi. Odpowiednio skonstruowany model przypisuje wyższe prawdopodobieństwo jednostkom, wśród których zaobserwowano wystąpienie zdarzenia niż takim, wśród których zdarzenia nie odnotowano.  
**- Gamma**

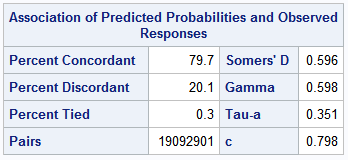
Miara Gamma podaje procentowy stosunek różnicy par dobrze i źle przyporządkowanych do ogólnej liczby par przyporządkowanych (bez tied). Stanowi ona proporcjonalną redukcję błędu, co oznacza, że przy zastosowaniu danego modelu (odpowiednich zmiennych objaśniających) mamy o tyle większe szanse na poprawne zaklasyfikowanie jednostek.

**- Sommersa**

Przedstawia stosunek par dobrze i źle przyporządkowanych do ogólnej liczby par (zarówno tych dobrze, jak i źle przyporządkowanych)

Statystyki D Somersa, Gamma oraz Tau-a testują niezależność zmiennych objaśnianej i objaśniających, na podstawie tablic kontyngencji. Statystyka gamma, dla skali porządkowej, jest nadwyżką zgodnych par ponad niezgodne wyrażaną jako procent w stosunku do wszystkich wyodrębnionych par (poza tied). Interpretować ją należy jako proporcjonalną redukcję błędu. Znając zmienne niezależne eliminujemy błąd oszacowania rangowania par o 59,8 %. Statystyka D Somersa jest modyfiakcją gamma i jest to nadwyżka par zgodnych wyrażona jako procent zgodnych, niezgodnych i związanych. Innymi słowy D Somers jest warunkowym prawdopodobieństwem, że para jest zgodna minus para jest niezgodna,.

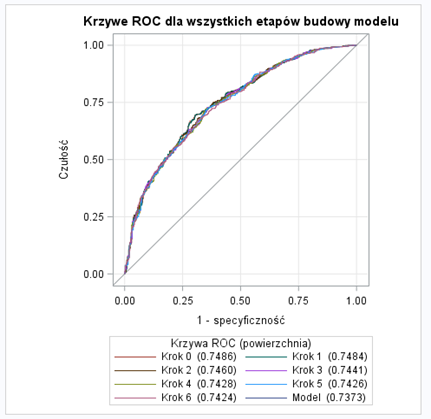
Statystyka c powstała jako zależność pomiędzy zgodnymi skojarzeniami par (zdolnością predykcyjną, dokładnością modelu) a popełnionymi przez model niezgonościami przyporządkowań (dokładniej: 1 – (źle sklasyfikowane poprawnie przyporządkowane pary/liczba niedopasowanych par). Im większa wartość c tym model lepiej przewiduje wartości zmiennej zależnej.



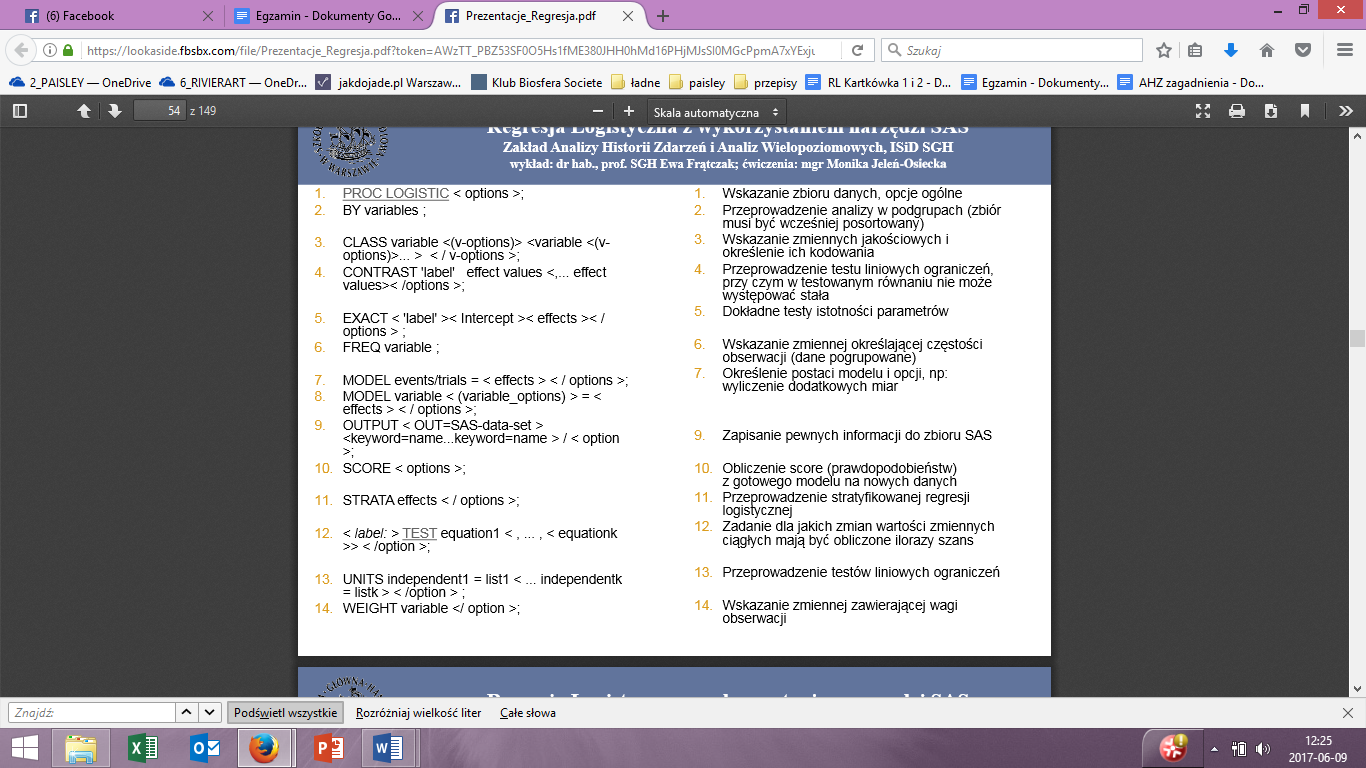
- **ROC** - umożliwia jednoczesną ocenę czułości i swoistości modelu ; im krzywa ROC jest bardziej wypukła, tym model lepszy (większe prawdopodobieństwo, że dobrze ustawimy próg odcięcia)

oś x-> 1-specificity = 1- TN/(TN+FP)

oś y -> sensitivity = TP/(TP+FN)



Pole pod krzywą ROC na poziomie 0,74 świadczy o dobrej klasyfikacji poziomów zmiennej objaśnianej oraz wysokiej trafności i dobroci modelu.



**PYTANIA ZAOCZNI 6.2017:**

**Regresja liniowa vs logistyczna**

**OR**

**RR**

**Znaleźć przedział ufność dla Odds ratio ma podstawie przedziału parametru, ocenić jego istotnośc**

**Opisać AIK i SC. Jak i do czego sie ich używa**

**Regresja wielomianowa vs porządkowa**

**Zinterpretować Odds dla regreji binarnej i uporządkowanej**

**Ocenić model na podstawie statystyk: gamma, sommers, tak kendalla, statystka c.**

**Ocenić test na kryterium proporcjonalnych szans**

**Napisać kody: binarnej, porządkowej, wielomianowej**

. Test Hosmera I Lemeshowa:to

testowanie różnic między przewidywaną a obserwowaną liczbą obserwacji w grupach. H0; model jest dobrze dopasowany do danych. jeśli p-value >0.05 to model jest dobrze dopasowany i odrzucamy alternatywną na rzecz hipotezy zerowej.

6.Pearson i Dewiancja

mówi nam jak dobrze model rozróżnia obiekty w obu grupach. wyższe pr dla tych które wystąpiło zdarzenie niż obiektom, gdzie nie wystąpiło. wtedy zbiory na siebie nie zachodzą. wszystko zależy od ilości unikatowych profil i estymowanych parametrów gdy ta różnica jest duża to nie interpretuje się. h0; model dobrze dopasowany do danych. bardzo duża liczba profili w stosunku do liczby obserwacji świadczy o tym że w wielu komórkach tablicy kontyngencji znalazły się pojedyncze jednostki schematu odpowiedzi na poszczególne pytania stanowiące zmienne objaśniające modelu. co ma wpływ na jakość predykcyjną

DLA MODELI wielomianu: może się okazać że liczba profili unikatowych jest zbyt duża i powoduje zakkłocenia bowiem może się okazać że wiele komórek tablicy kontygencji będzie pusta. (nierównomierny rozkad obserwacji zniekształca analizę)