Politechnika Krakowska
Wydział Informatyki i Telekomunikacji

Sprawozdanie z przedmiotu:

Statystyka i Probabilistyka

Projekt nr 3

Temat:

Model regresji logistycznej prostej i wielokrotnej z możliwością włączenia zmiennych niezależnych typu jakościowego

Wykonał: **Rafał Gęgotek**
Kierunek: Informatyka
Stopień studiów: II stopnia
Specjalizacja: Data Science
Rok akademicki: 2020/2021

1. Cel projekty

Celem projektu jest zastosowanie trzech modeli regresji logistycznej. W pierwszym etapie należy użyć regresji ze zmienną ilościową, następnie dla tej samej zmiennej objaśnianej użyć zmiennej jakościowej, a na koniec zastosować regresję wieloraką z uwzględnieniem wcześniejszych zmiennych niezależnych oraz dwóch dodatkowych zmiennych ilościowych.

W ramach projektu należy znaleźć odpowiedni zestaw danych, który zostanie poddany analizie wykonanych na nim regresji logistycznej, w oparciu o techniki poznane na zajęciach projektowych i wyciągnięciu odpowiednich wniosków.

2. Zbiór danych

Badany zestaw danych dotyczą upublicznionych statystyk zebranych z 4 różnych placówek leczniczo-badawczych jakimi są:

- Węgierski Instytut Kardiologii. Budapeszt
- Szpital Uniwersytecki, Zurych, Szwajcaria
- Szpital Uniwersytecki, Bazylea, Szwajcaria
- VA Medical Center, Long Beach and Cleveland Clinic Foundation

Statystyki te zawierają informację na temat parametrów sercowo-naczyniowych w odniesieniu, czy dany pacjent ma zdiagnozowaną chorobę serca. Oryginalny zestaw danych składa się 76 atrybutów, natomiast zredukowany z 14 i łącznie liczy 303 instancje. Zdjęcie poglądowe poniżej.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
12	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
13	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
14	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
15	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1

Rysunek 1. Wycinek tabeli badanego zestawu danych

Poszczególne parametry oznaczają odpowiednio:

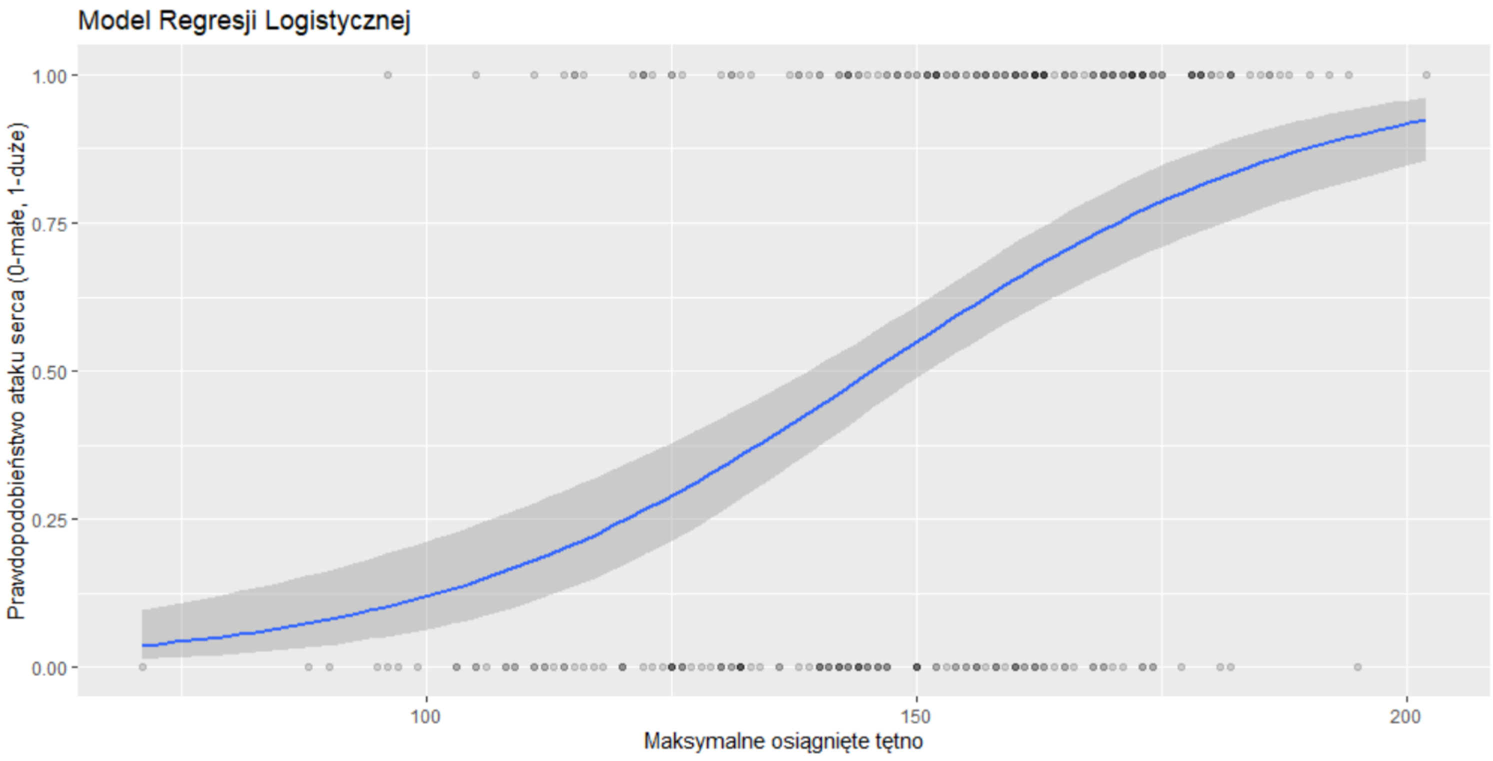
- age - wiek,
- sex – płeć (1 – mężczyzna, 2 – kobieta)
- cp - rodzaj bólu w klatce piersiowej (4 wartości),
- **trestbps** - spoczynkowe ciśnienie krwi,
- **chol** – poziom cholesterol w mg/dl,
- fbs - poziom cukru we krwi na czczo (1 dla większego niż 120 mg/dl),
- restecg - spoczynkowe wyniki elektrokardiograficzne,
- **thalach** –maksymalne osiągnięte tętno,
- **exang** - dławica piersiowa wywołana wysiłkiem fizycznym,
- oldpeak - obniżenie odcinka ST wywołane wysiłkiem fizycznym,
- slope - nachylenie szczytowego odcinka ST podczas ćwiczenia,
- ca - liczba głównych naczyń (0-3) pokolorowanych fluoroskopią,
- **target** – szansa na wystąpienia zawału serca (0 - małe, 1 - duże)

Głównym celem badanego zbioru jest zbadanie zależności prawdopodobieństwa wystąpienia zawału serca, a zmiennymi niezależnymi. W pierwszym etapie zmienną ilościową będzie maksymalnie osiągnięte tętno przez pacjenta. W następnym modelu regresji zmienną jakościową będzie dławica piersiowa wywołana wysiłkiem fizycznym. Na koniec w modelu regresji wielokrotnej zostaną dodane do obu wspomnianych parametrów, zmienne spoczynkowego ciśnienia krwi oraz poziom cholesterolu.

Główną przyczyna wybrania takiego zbioru i parametrów jest chęć zbadania w jakim stopniu, na podstawie zmiennych niezależnych można określić osoby mające większą skłonności do wystąpienia w przyszłości zawału serca.

3. Regresja liniowa ze zmienna niezależną ilościową

W pierwszym modelu regresji prawdopodobieństwo zawału serca będzie objaśniane przy pomocy zmiennej maksymalnie osiągniętego tętna. Poniżej została zwizualizowana krzywa prawdopodobieństwa tego modelu, z której możemy zaobserwować pewną zależność, świadcząco o większym ryzyku zawału serca dla wysokich wskazań tętna badanych osób.



Rysunek 2. Wizualizacja modelu regresji logistycznej dla zmiennej ilościowej maksymalnego tętna

W pierwszym kroku diagnostyki, należy poddać analizie podsumowanie modelu. W części dotyczącej odchyleń reszt możemy odczytać, iż odchylenia składników resztowych są w normie mieszczącej się w skali <-3, 3>.

Kolejną ważną kwestią jest istotność zmiennych niezależnych, gdzie biorąc pod uwagę parametr p, możemy zdecydowanie odrzucić hipotezę zerową. Otóż zarówno dla zmiennej ilościowej 'thalach" jak i wyrazu wolnego wartość p jest znacznie mniejsza niż 5%, a więc zmienne są istotne, co jeszcze podkreślają znajdujące się przy zmiennych 3 gwiazdki.

Ponadto dla wskazań zmiennej jakościowej można odczytać, iż zwiększenie się o jedną jednostkę (uderzenie na minutę) maksymalnego tętna wiąże się ze wzrostem logarytmicznym prawdopodobieństwo odnoszącego się do wystąpienia zawału serca o 0.043951. Tak więc z każdym wzrostem o jedno uderzenia na minutę tętna, szansa na wystąpienie zawału serca wzrasta o współczynnik 1.0043951.

Dodatkowo jakość modelu określa też wskaźnik AIC (Kryterium informacyjne Akaikego), którego mniejsze wskazania oznaczają, że model jest bliższy prawdy. W tym przypadku wynosi on 363.26 i będzie brany pod uwagę podczas porównywania z innymi analizowanymi modelami.

```
> model1 <- glm(target ~ thalach, family = "binomial", data = data)
> summary(model1)

Call:
glm(formula = target ~ thalach, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1383  -1.0780   0.6043   0.9200   2.1354

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.391452   0.987133  -6.475 9.50e-11 ***
thalach      0.043951   0.006531   6.729 1.71e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 359.26  on 301  degrees of freedom
AIC: 363.26

Number of Fisher Scoring iterations: 4

> exp(coef(model1))
(Intercept)     thalach
0.001675821 1.044931450
```

Rysunek 3. Podsumowanie modelu regresji logistycznej dla zmiennej ilościowej

Pomimo, iż w regresji logistycznej nie ma możliwości stosowania współczynnika R^2, to jednak używa się pseudo R kwadrat MCFadden'a, w którym zamiast stosować metodę najmniejszych kwadratów, stosuje się metodę największej wiarygodności. Tak jak pokazana na poniższym zdjęciu wartość tego współczynnika dla modelu, ze zmienna ilościową wynosi 0.13978, co można uznać za przeciętny wynik.

```
> pscl::pR2(model1)["McFadden"]
fitting null model for pseudo-r2
 McFadden
0.1397888
```

Rysunek 4 Wskazanie pseudo R^2 McFadden'a dla modelu ze zmienna ilościową

W kolejnym kroku można przedstawić jak prezentują się przedziały ufności dla zmiennych niezależnych. Biorąc pod uwagę prawdopodobieństwa 95%, przedziały dla zmiennej ilościowej mieszczą się w zakresie od 0.031 do 0.057.

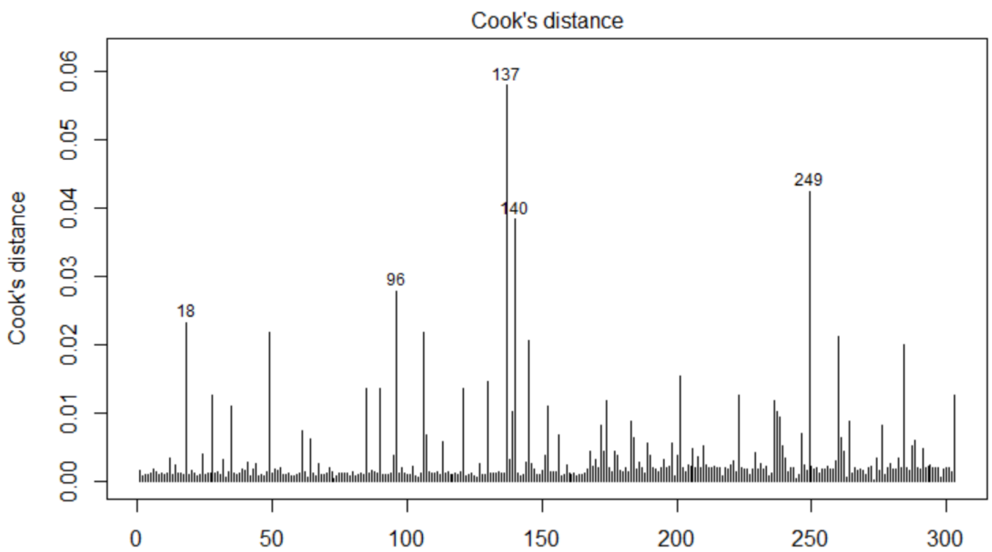
```
> confint(model1)
Waiting for profiling to be done...
                  2.5 %      97.5 %
(Intercept) -8.41257396 -4.53271457
thalach      0.03164292  0.05731192
```

Rysunek 5 Przedziały ufności dla modelu regresji logistycznej, dla zmiennej ilościowej

Dla standardowych składników resztowych, w modelach regresji logistycznej nie trzeba poddawać analizy homoskedastyczności i korelacji reszt. Jedyną diagnostyką w tym przypadku jest sprawdzenie wartości odstających. Tak jak wcześniej zostało to opisane, przy analizie podsumowania modelu nie było zidentyfikowanych takich wartości. Dla pewności możemy jeszcze się upewnić sprawdzając standardowe reszty, tak jak pokazano to poniżej, gdzie żadna wartość nie przekracza granicznych.

```
> model1_data <- augment(model1) %>% mutate(index = 1:n())
> model1_data %>% filter(abs(.std.resid) > 3)
# A tibble: 0 x 9
# ... with 9 variables: target <int>, thalach <int>, .fitted <dbl>, .resid <dbl>,
#   .std.resid <dbl>, .hat <dbl>, .sigma <dbl>, .cooksd <dbl>, index <int>
```

Rysunek 6 Identyfikacja obserwacji odstających dla modelu regresji logistycznej, ze zmienną ilościową



Rysunek 7 Wizualizacja obserwacji najbardziej odstających dla modelu regresji logistycznej, ze zmienną ilościową

Dodatkowo niezależnie od tego, czy są zidentyfikowane obserwacje odstające, możemy zwizualizować te wartości które najbardziej odbiegają od reszty. Na powyższym wykresie dystansu Cook'a zostało oznaczonych 5 takich wartości.

Mając gotowy model regresji, w łatwy sposób możemy obliczyć prognozę prawdopodobieństwa dla zmiennej zależnej.

```
> predict(model1, data.frame(thalach = c(100, 130, 160)), type = "response")
        1         2         3 
0.1195895 0.3367559 0.6549238 
```

Rysunek 8 Predykcja modelu regresji logistycznej, dla zmiennej ilościowej

Tak więc dla osób , których maksymalnie osiągnę tętno wynosi 160 uderzeń na minutę to prawdopodobieństwo wystąpienie zawału serca wynosi aż 65%. Natomiast dla wskazań maksymalnego tętna o wartości 130 i 100 jest to odpowiednio 33% i 11%.

Na podstawie otrzymanych wyników dla zastosowanego zestawu danych, możemy wyciągnąć wniosek, iż istnieje zależność liniowa pomiędzy prawdopodobieństwem zawału serca, a maksymalnym tętnem. Sam model nie posiada oznak wartości odstających, natomiast biorąc pod uwagę współczynnik pseudo R kwadrat możemy stwierdzić, że model nie jest jednak doskonały.

Podsumowując tą cześć diagnostyczną, im większe maksymalne tętno, prawdopodobieństwo wystąpienia zawału wzrasta. W taki wypadku dużo bardziej narażone są osoby mające skłonności do zaburzeniem pracy serca, jakim jest tachykardia (wysokie tętno).

4. Regresja liniowa ze zmienna niezależną jakościową

W tej części modelu regresji prawdopodobieństwo zawału serca będzie objaśniane przy pomocy zmiennej jakościowej dotyczącej dławicy piersiowej wywołana wysiłkiem fizycznym (1 – występowanie zaburzenia, 0 – brak zaburzenia).

Proces diagnostyki modelu jest podobny jak w poprzednim punkcie. W pierwszym kroku analizujemy podsumowanie modelu.

```
> model2 <- glm(target ~ exang, family = "binomial", data = data)
> summary(model2)

Call:
glm(formula = target ~ exang, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5434  -0.7272   0.8512   0.8512   1.7086

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8287     0.1522   5.444 5.21e-08 ***
exang        -2.0239     0.2825  -7.164 7.82e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 357.90  on 301  degrees of freedom
AIC: 361.9

Number of Fisher Scoring iterations: 4

> exp(coef(model2))
(Intercept)       exang
  2.2903226   0.1321349
```

Rysunek 9 Podsumowanie modelu regresji logistycznej ze zmienna jakościową

Odnośnie zmiennych niezależnych, biorąc pod uwagę parametr p, możemy zdecydowanie odrzucić hipotezę zerową. Zarówno dla zmiennej ilościowej ‘exang” jak i wyrazu wolnego. Wartość p jest zauważalnie mniejsza niż 5%, a więc zmienne są istotne. Podobnie jak w poprzednim modelu każda zmienna jest oznaczona również 3 gwiazdkami, świadczącymi o istotności na podstawie parametru p.

W części dotyczącej odchyleń reszt możemy odczytać, iż odchylenia składników resztowych są w normie mieszczącej się w skali <-3, 3>. Tak więc ponownie nie będzie żadnych wartości odstających

Dodatkowo jakość modelu określona przez wskaźnik AIC jest równa 361.9. Odwołując się do poprzedniego modelu można stwierdzić, że na podstawie tego parametru oba modele podobnie odwzorowują zmienna objaśnianą. Wartość ta wypada na korzyść zmiennej ‘exang’, lecz tylko o 1.3 jednostki.

Podobnie możemy to przedstawić przy pomocy metody anova, wykorzystującej test ‘chi kwadrat’.

```
> anova(model1, model2, test = "Chisq")
Analysis of Deviance Table

Model 1: target ~ thalach
Model 2: target ~ exang
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       301     359.26
2       301     357.90  0    1.354
```

Rysunek 10. Porównanie modeli regresji logistycznej ze zmienna jakościową i ilościową przy pomocy testu ‘chi kwadrat’

Dla modelu w którym obie zmienne są jakościowe, możemy obliczyć predykcję dla występowania poszczególnych wartości.

```
> predict(model2, data.frame(exang = c(1, 0)), type = "response")
        1         2
0.2323232 0.6960784
```

Rysunek 11. Predykcja modelu regresji logistycznej, dla zmiennej ilościowej

Tak jak pokazano powyżej, dla osób cierpiących na zaburzenie dławicy piersiowej podczas wysiłku fizycznego, prawdopodobieństwo zawału serca jest równe 23%. Natomiast w przeciwnym wypadku prawdopodobieństwo to wynosi aż 69%.

Biorąc pod uwagę kolejny parametr diagnostyczny jakim jest pseudo R kwadrat MCFadden’a, można stwierdzić, iż model z tą konkretną zmienna jakościową jest lepszy od modelu ze zmienna ilościową. Współczynnik ten wynosi 0.14303. Tak więc ponownie należy przyznać, iż model słabo odwzoruje zmienną jakościową.

```
> pscl::pR2(model2)["McFadden"]
fitting null model for pseudo-r2
 McFadden
0.1430308
>
```

Rysunek 12. Wskazanie pseudo R^2 McFadden'a dla modelu ze zmienna jakościową

Na podstawie otrzymanych wyników, możemy wyciągnąć wniosek, iż istnieje zależność liniowa pomiędzy prawdopodobieństwem zawału serca, a zmienna jakościową dotyczącą dławicy piersiowej wywołanej wysiłkiem fizycznym. Biorąc jednak pod uwagę współczynnik pseudo R kwadrat możemy stwierdzić, że model jest słaby (pomimo nieco lepszych wyników od poprzedniego modelu).

5. Regresja wielokrotna

Dla ostatniego przykładu regresji zmienna zawału serca będzie objaśniana przy pomocy dwóch wcześniej omawianych (maksymalne tętno i dławica piersiowa podczas wysiłku fizycznego) oraz dodatkowo dwóch kolejnych zmiennych ilościowych, jakimi są: poziom cholesterolu i ciśnienie krwi.

```
> model3 <- glm(target ~ exang + chol + trestbps + thalach, family = "binomial", data = data)
> summary(model3)

Call:
glm(formula = target ~ exang + chol + trestbps + thalach, family = "binomial",
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1206  -0.7712   0.5046   0.8186   2.2345

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.552589   1.533831  -1.012   0.3114
exang       -1.586804   0.304305  -5.215 1.84e-07 ***
chol        -0.002841   0.002579  -1.102   0.2706
trestbps    -0.016646   0.008030  -2.073   0.0382 *
thalach      0.034351   0.006827   5.032 4.86e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 322.99  on 298  degrees of freedom
AIC: 332.99

Number of Fisher Scoring iterations: 4
```

Rysunek 13. Podsumowanie modelu regresji logistycznej wielokrotnej

Na podstawie otrzymanych wyników, można stwierdzić, iż model ten najlepiej odwzorowuje zmienną objaśnianą. Świadczy o tym chociażby współczynnik AIC, mający wartość 332.99.

Biorąc pod uwagę odchylenia reszt możemy odczytać, iż ponownie odchylenia składników resztowych mieszczą się w normie w przedziale <-3, 3>.

Analizując zmienne niezależne, należy przyznać, iż dwie zmienne są bardzo istotne, a są nimi ‘thalach’ oraz ‘exang’, a więc dwie wcześniej już analizowane osobno. Dodatkowo zmienna ciśnienia krwi charakteryzuje się w modelu wartością p mniejszą od 5%, tak że została oznaczona jedną gwiazdką istotności. Z pośród wszystkich zmiennych najgorzej wypadł poziom cholesterolu, który jest nie istotny dla danego modelu.

Na tej podstawi można by pokusić się o zredukowanie zmiennych niezależnych o najmniej istotną, i sprawdzeniu czy nie pogorszy się jego skuteczność. Natomiast nie to jest istotą projektu, dlatego też końcowy wzór opisujący prawdopodobieństwo wystąpienie zawału serca dla zastosowania 4 zmiennych, prezentuje się następująco:

$$p(X) = \frac{e^{-1.552\ -\ 1.5868*X1-0.0028*X2-0.0166*X3+0.0343*X4}}{1 + e^{-1.552\ -\ 1.5868*X1-0.0028*X2-0.0166*X3+0.0343*X4}}$$

gdzie:

x1 – exang – dławica piersiowa wywołana wysiłkiem fizycznym

x2 – chol – poziom cholesterolu

x3 – trestpbs – ciśnienie krwi

x4 – thalach – maksymalnie osiągnięte tętno

Biorąc pod uwagę wszystkie przedstawione modele, możemy je porównać na podstawie wskazań pseudo R kwadrat McFadden’a.

W takim wypadku modele jeden i dwa, a więc wcześniej omawiane, maja podobne wartości tego parametru na poziomie 14 %. Natomiast znacznie uległ poprawie model regresji uwzględnieniu w nim 4 zmiennych, tak że teraz poziom jego pseudo R kwadrat wynosi 22.66%.

Nie jest to oszałamiający wynik, jednakże należy przyznać, iż zastosowanie większej liczby zmiennych niezależnych dało zauważalnie lepszy efekt.

```
> list(model1 = pscl::pR2(model1)["McFadden"],
+      model2 = pscl::pR2(model2)["McFadden"],
+      model3 = pscl::pR2(model3)["McFadden"])
fitting null model for pseudo-r2
fitting null model for pseudo-r2
fitting null model for pseudo-r2
$model1
 McFadden
0.1397888

$model2
 McFadden
0.1430308

$model3
 McFadden
0.2266198
```

Rysunek 14. Wskazanie pseudo R^2 McFadden'a dla trzech omawianych modeli

Podobne porównanie możemy przeprowadzić przy pomocy metody anova i testu 'chi kwadrat'.

```
> anova(model1, model3, test = "Chisq")
Analysis of Deviance Table

Model 1: target ~ thalach
Model 2: target ~ exang + chol + trestbps + thalach
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       301     359.26
2       298     322.99  3   36.264 6.585e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(model2, model3, test = "Chisq")
Analysis of Deviance Table

Model 1: target ~ exang
Model 2: target ~ exang + chol + trestbps + thalach
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       301     357.90
2       298     322.99  3    34.91 1.273e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Rysunek 15. Porównanie modeli przy pomocy testu 'chi kwadrat'

Można zdecydowanie stwierdzić, iż dla obu pokazanych przykładów tego testu, model regresji wielokrotnej jest bardziej istotny od modeli regresji logistycznej z jedną zmienną, które były omawiane wcześniej. Dla obu przypadków wartość p jest znacznie mniejsza, niż 5%. Również należy wspomnieć o charakterystycznych 3 gwiazdkach, świadczących o istotności modelu dla regresji logistycznej wielokrotnej.

Biorąc pod uwagę standardowe składniki resztowe, możemy upewnić się czy nie ma w modelu wartości odstających. W celu lepszej wizualizacji zamiast zakresu -3 do 3 zostanie użyty zakres -2 do 2, aby wyświetlić przynajmniej wartości najbliżej graniczne.

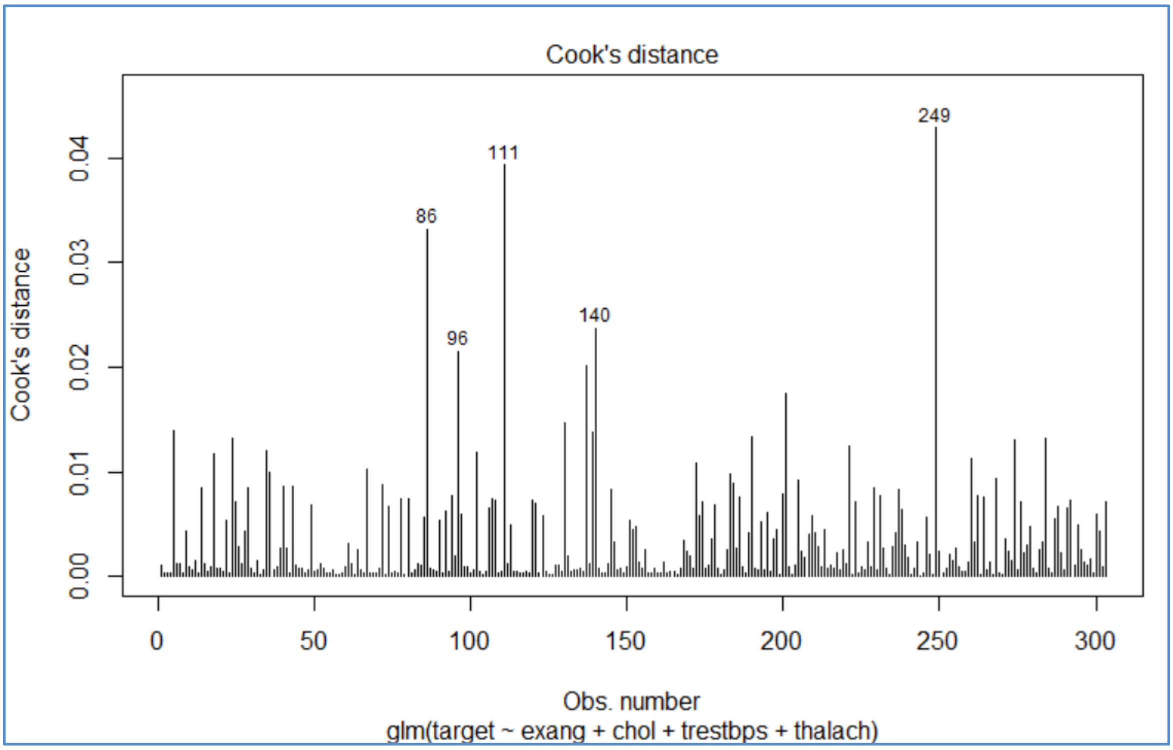
```
> model3_data %>% filter(abs(.std.resid) > 2)
# A tibble: 3 x 12
  target exang  chol trestbps thalach .fitted .resid .std.resid   .hat .sigma .cooksd index
   <int> <int> <int>    <int>   <int>   <dbl>  <dbl>      <dbl>  <dbl>  <dbl>   <dbl> <int>
1      1     1   226      142     111   -2.33   2.20       2.21 0.0102   1.03  0.0215    96
2      1     1   263      128     105   -2.41   2.23       2.25 0.0104   1.03  0.0237   140
3      0     0   197      110     177    2.14  -2.12      -2.13 0.0101   1.04  0.0174   201
```

Rysunek 16. Identyfikacja obserwacji odstających dla modelu regresji wielokrotnej (przedział od -2 do 2)

Tak jak pokazano są tylko trzy wartości które, przekraczają podany zakres, jednak nie są to wartości odstające przekraczające abs(3). Dodatkowo możemy przedstawić graficznie wartości będące najbardziej odstające w zbioru (rys. 18), oraz krótkie podsumowanie dotyczące wskazanych punktów (rys. 17).

```
> model3_data %>% top_n(5, .cooksd)
# A tibble: 5 x 12
  target exang  chol trestbps thalach .fitted .resid
   <int> <int> <int>    <int>   <int>   <dbl>  <dbl>
1      1     0   564      115     160   0.427   1.00
2      1     1   226      142     111  -2.33    2.20
3      1     1   325      180     154  -1.77    1.96
4      1     1   263      128     105  -2.41    2.23
5      0     0   283      192     195   1.15   -1.69
# ... with 5 more variables: .std.resid <dbl>,
#   .hat <dbl>, .sigma <dbl>, .cooksd <dbl>,
#   index <int>
> 
```

Rysunek 17. Analiza 5 największych wartości odstających dla modelu regresji logistycznej wielokrotnej



Rysunek 18. Wykres dystansu Cook'a dla modelu regresji logistycznej wielokrotnej, z uwzględnieniem 5 największych wartości odstających

6. Wnioski

Na podstawie otrzymanych wyników dla zastosowanego zestawu danych, możemy wyciągnąć wniosek, iż istnieje zależność liniowa pomiędzy prawdopodobieństwem zawału serca, a maksymalnym osiąganym tętnem, dławica piersiowa wywołaną wysiłkiem fizycznego, oraz w mniejszym stopniu ciśnieniem krwi. Końcowy model regresji wielokrotnej uwzględniający te zmienne oraz poziom cholesterolu (zmienna nieistotna na podstawie parametru p) wskazuje na wartość objaśnianą współczynnikiem pseudo R kwadrat na poziomie 22%, co można zaliczyć jako przeciętny wynik.

Jednakże uwzględnić trzeba, że dla dwóch pierwszych analizowanych modeli regresji z jedną zmienną niezależną, wartość ta była znacznie mniejsza, dlatego też model regresji wielokrotnej jest bardziej istotny.

Podsumowując, im większe maksymalne tętno oraz brak występowania dławicy piersiowej podczas wysiłku, a także mniejsze ciśnienie krwi, tym większe jest prawdopodobieństwo wystąpienia zawału serca. Jednak na podstawie informacje o jakości otrzymanego modelu, należy zaznaczyć, że uwzględnione zmienne niezależne nie odzwierciedlają w dobrym stopniu zmiennej objaśnianej.