Diagnostyka pacjenta z cukrzycą

Yauheni Semianiuk

Warszawa 2023

Spis treści

Wstęp

Diabetes mellitus, znana w Polsce pod nazwą cukrzyca, jest grupą powszechnych chorób endokrynologicznych charakteryzujących się utrzymującym się wysokim poziomem cukru we krwi. Cukrzyca jest spowodowana albo niedostatecznym wydzieleniu insuliny w trzustce z powodu autoimmunologicznego, albo zaburzeniem równowagi między poziomem cukru we krwi a insuliną, i może być przyspieszona przez ciążę. Deklaracja St Vincent przyjęta w 1989 roku była wynikiem międzynarodowych starań zmierzających do polepszenia opieki nad chorymi na cukrzycę. Takie działania mają znaczenie zarówno w perspektywie indywidualnej jakości życia, jak i ekonomiczne – wydatki na cukrzycę okazały się główną przyczyną obciążającą systemy opieki zdrowotnej. Globalne straty spowodowane cukrzycą szacowane są w 2015 na 1,31 bilionów dolarów amerykańskich. Stanowi to 1,8% globalnego produktu krajowego brutto.[1]

Zastosowanie podejścia bayesowskiego wydaje się szczególnie istotnym w badaniach medycznych. Temu służy kilka powodów. Po pierwsze, istnieją tysięcy prac związanych z badaniem różnych chorób ludzkich. Ponieważ cukrzyca jest jedną z najbardziej znanych chorób XXI wieku, łatwo można znaleźć dziesiątki gotowych rozważań zarówno teoretycznych, jak i praktycznych. Po drugie, medycyna przez długi czas była (i jest) nauką ekspercko-orientowaną. Oznacza to, że niemałoważne (jeżeli nie kluczowe) jest doświadczenie zdobyte przez lekarze w praktyce. Pozwała to na uzyskanie rozkładów parametrów a priori. Celem niniejszej pracy jest ocena czynników wpływających na (nie)występowanie cukrzycy w oparciu o zdobytą wiedzą a priori i zebrane dane.

[1] https://diabetesjournals.org/care/article/41/5/963/36522/Global-Economic-Burden-of-Diabetes-in-Adults , dostęp 07.02.2023

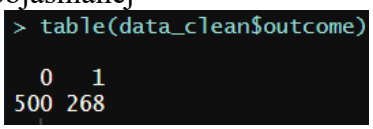
Rozdział I: Opis i analiza danych

Do analizy wykorzystany zbiór danych, przygotowany przez Narodowy Instytut Cukrzycy i Chorób Trawiennych i Nerkowych w Stanach Zjednoczonych. Zbiór zawiera 768 obserwacji zebranych w maju 1990 roku. Przedstawione są również następujące zmienne objaśniane:

- pregnancies – liczba ciąż w trakcie życia,
- glucose – stężenie glukozy,
- blood_pressure –ciśnienie krwi (mm Hg),
- skin_thickness – grubość fałdu skórnego tricepsa (mm),
- insulin – insulina w surowicy (mu U/ml),
- bmi – wskaźnik masy ciała (waga w kg/(wzrost w m)^2),
- diabetes_pedigree_func – funkcja rodowodu cukrzycy,
- age – wiek w latach.

Analiza zaczyna się od czyszczenia danych. Po pierwsze, przedstawione dane nie są zbilansowane, tzn. występuje znacznie więcej osób nie chorych na cukrzycę (rysunek 1). Próba została skrócona metodą undersampling w ten sposób, aby zostało po 268 osób w każdej kategorii zmiennej objaśnianej. Po drugie, w celach przestrzegania zasad bieżącej pracy, każda z grup została obcięta o losowo wybrane 18 obserwacji. W ten sposób otrzymana została końcowa ramka danych z 500 wierszami.

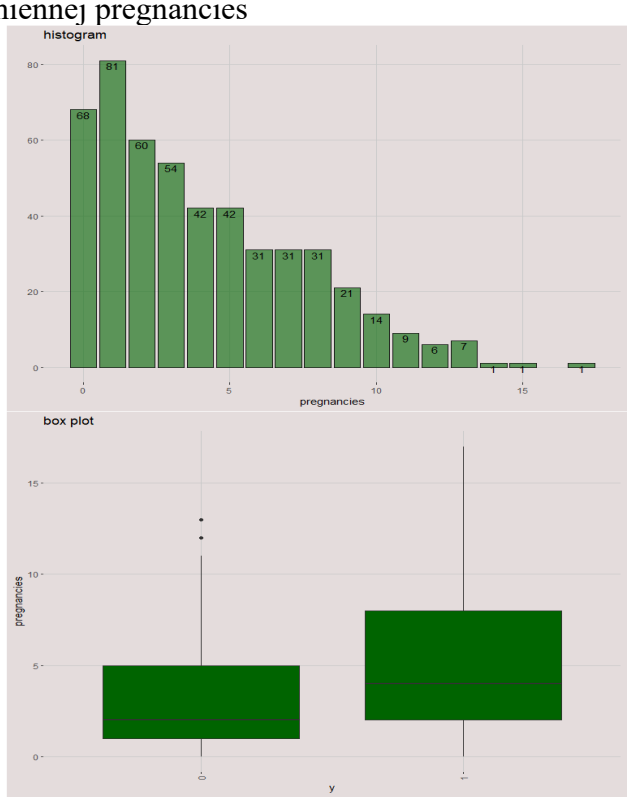
Rysunek 1. Rozkład zmiennej objaśnianej



Źródło: opracowanie własne

Po ucinaniu zbioru przeanalizowane zostały przedstawione zmienne objaśniające.

Rysunek 2. Rozkład zmiennej pregnancies

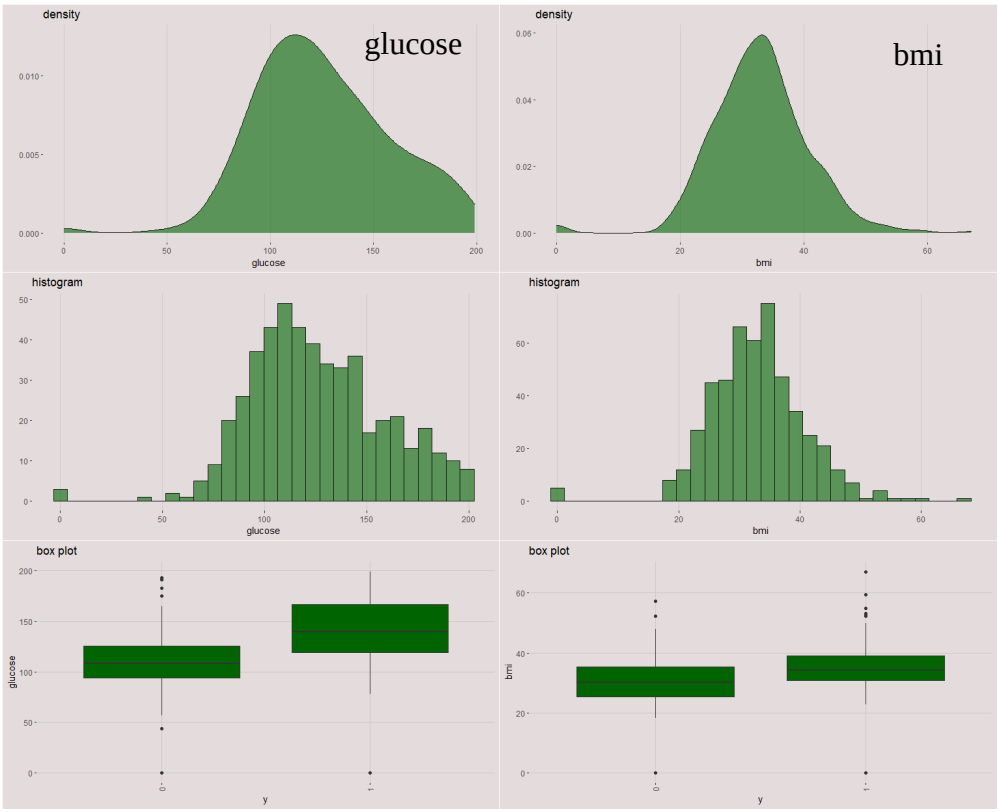


Źródło: opracowanie własne

Pierwszą jest zmienna numeryczna pregnancies z prawoskośnym rozkładem o maksymalnej wartości równej 17. Jak wynika z wykresu 1.1, liczba ciąży ma istotny związek statystyczny z występowaniem u pacjenta cukrzycy, ponieważ mediana w tym przypadku jest większa.

Następnie przeanalizowane zmienne glucose (po lewiej stronie) i bmi (po prawej stronie). Mają oni dość scentralizowane rozkłady o jednej dominancie. Obie zmienne mają jawną zależność statystyczną ze zmiennej objaśnianą y ze względu na różne poziomy pudełek. Zarówno jak i dla zmiennej pregnancies, zakłada się że będą one miały mocny wpływ na występowanie cukrzycy.

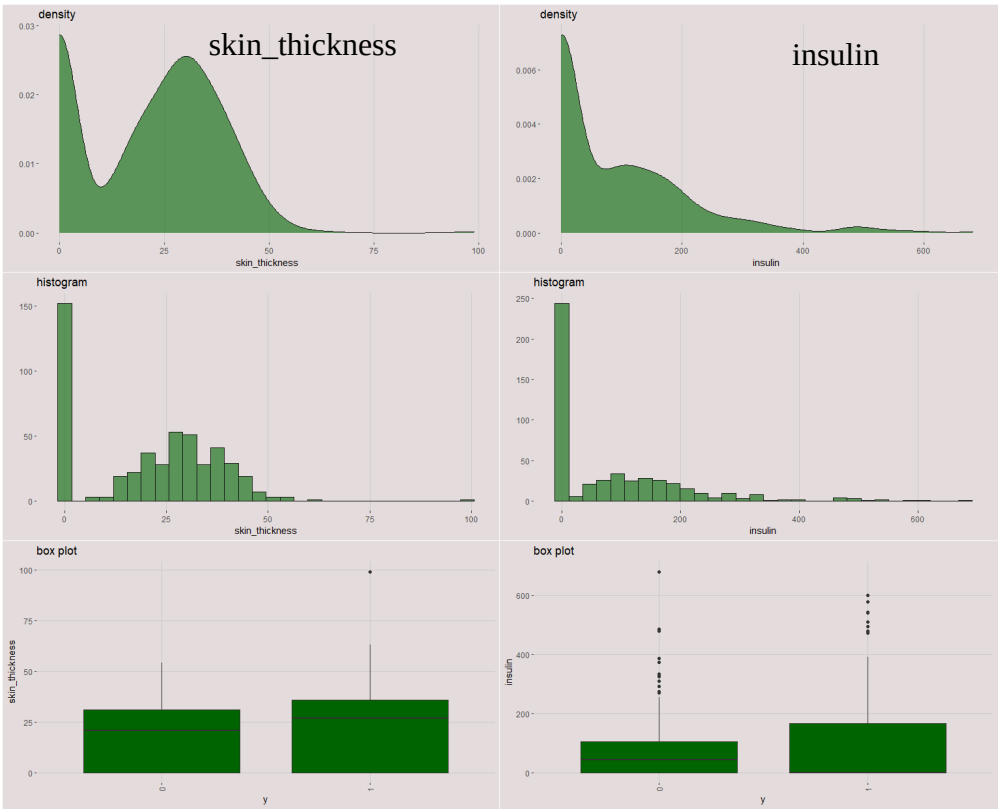
Rysunek 3. Rozkłady zmiennych glucose i bmi



Źródło: opracowanie własne

Dalsza analiza dotyczyła poziomu insuliny oraz grubości skóry. Jak wynika z rysunku 3, zmienne insulin oraz skin_thickness są bardzo do siebie podobne, co może świadczyć o dużej korelacji zmiennych. Co więcej, widzimy mocną kumulację wokół zera. W bieżącym zbiorze danych zera oznaczają wartości nieznane. W końcu wykresy pudełkowe są zbliżone dla obu klasów zmiennej objaśnianej. Powyższe fakty przekonują do decyzji o nieuwzględnieniu tych zmiennych na kolejnych etapach naszej pracy, zamiast skrócenia zbioru o 244 (w przypadku insuliny) lub 152 obserwacje (w przypadku grubości skóry).

Rysunek 4. Rozkłady zmiennych skin thickness oraz insulin



Źródło: opracowanie własne

Nie ma też powodów liczyć sam fakt występowania nieznanych wartości zmiennych za jakąkolwiek zależność statystyczną. Wskazują na to testy χ, oparte na łącznych rozkładach zmiennych objaśniających i zmiennej objaśnianej (rysunek 5).

Rysunek 5. Testy χ zmiennych skin_thickness oraz insulin

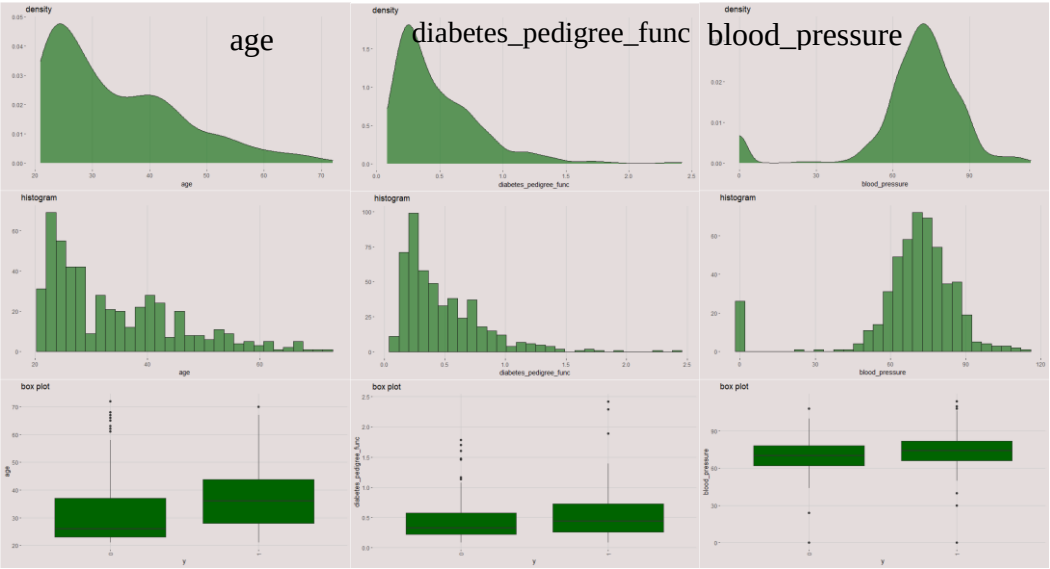
```
> table(chi_df$y, chi_df$insulin)

     0   1
  0 133 117
  1 123 127
> chisq.test(chi_df$y, chi_df$insulin)

        Pearson's Chi-squared test with Yates' continuity correction

data:  chi_df$y and chi_df$insulin
X-squared = 0.64837, df = 1, p-value = 0.4207

> table(chi_df$y, chi_df$skin_thickness)

     0   1
  0 179  71
  1 169  81
> chisq.test(chi_df$y, chi_df$skin_thickness)

        Pearson's Chi-squared test with Yates' continuity correction

data:  chi_df$y and chi_df$skin_thickness
X-squared = 0.76565, df = 1, p-value = 0.3816
```

Źródło: opracowanie własne

Następnymi z kolei są zmienne age i diabetes_pedigree_func. Mają oni mocny prawoskośny rozkład o wąskim ogonie oraz istotną różnice w medianie dla różnych wartości

zmiennej objaśnianej. Oznacza to, że wiek i genealogia mają wpływ na to czy pacjent będzie chory na cukrzycę.

Rysunek 6. Rozkład zmiennych j age, diabetes_pedigree_func i blood_pressure.



Źródło: opracowanie własne

Na koniec analizie uległa zmienna blood_pressure (rysunek 6). Jej rozkład przypomina rozkład zmiennej bmi. Jednak w odróżnieniu od zmiennej bmi, wpływ na zmienną objaśnianą już nie jest taki oczywisty. Co więcej, 26 wartości ciśnienia są nieznane (chyba że to są martwi ludzie), więc obserwacje te zostały usunięte z pierwotnego zbioru.

Po uwzględnieniu informacji zdobytych podczas wstępnej analizy danych oceniona została korelacja zachodząca pomiędzy zmiennymi.

Rysunek 7. Korelacja zmiennych

variable	y	pregnancies	glucose	blood_pressure	bmi	diabetes_pedigree_func	age
<chr>	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*
y	1	0.24	0.47	0.2	0.29	0.18	0.29
pregnancies	0.24	1	0.11	0.19	0.01	-0.05	0.58
glucose	0.47	0.11	1	0.25	0.23	0.11	0.22
blood_pressure	0.2	0.19	0.25	1	0.33	0	0.35
bmi	0.29	0.01	0.23	0.33	1	0.16	0.01
diabetes_pedigree_func	0.18	-0.05	0.11	0	0.16	1	0.02
age	0.29	0.58	0.22	0.35	0.01	0.02	1

Źródło: opracowanie własne

W wyniku okazało się, że najbardziej istotną zmienną jest zmienna glucose. Co więcej nie zauważalne jest zjawisko współliniowości.

Rozdział II: Model logit

Kolejna część pracy dotyczyła zbudowania modelu logitowego. Jednak w modelu, oszacowanym na podstawie danych z rozdziału I, nieistotną okazała się zmienna blood_pressure. W związku z tym, że to się zgadzało z wynikami eksploracyjnej analizy danych, była podjęta decyzja o usunięciu tej zmiennej z modelu. Ostateczny model można opisać za pomocą następującej formuły:

$$\mathrm{g}(\mathrm{E}(\mathrm{Y})) = \mu(x) = -8.63 - 0.10 \times pregnancies + 0.03 \times glucose + 0.07 \times BMI + 1.10 \times DPF + 0.03 \times age,$$

$$g(\cdot) = ln\frac{p(x)}{1-p(x)}$$

Rysunek 8. Model logit

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.625390   0.876843  -9.837  < 2e-16 ***
pregnancies              0.102133   0.041890   2.438  0.01477 *
glucose                  0.033242   0.004272   7.782 7.16e-15 ***
bmi                      0.073936   0.017032   4.341 1.42e-05 ***
diabetes_pedigree_func   1.099568   0.358247   3.069  0.00215 **
age                      0.029931   0.012424   2.409  0.01599 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 657.07  on 473  degrees of freedom
Residual deviance: 475.15  on 468  degrees of freedom
AIC: 487.15
```

Źródło: opracowanie własne

Wpływ pozostałych zmiennych objaśniających był zakwalifikowany jako istotny.

Rozdział III: Informacja a priori

Na podstawie podobnych pracy w temacie modelowania zachorowania na cukrzycę zostały wybrane rozkłady oraz parametry rozkładów a priori dotyczące parametrów modelu logitowego. Są to rozkłady normalne o średnich i wariancjach przedstawionych poniżej:

Tabela 1. Wartości parametrów a priori

Zmienna	Średnia	Wariancja
(Intercept)	-20.76	46.63
pregnancies	0.33	0.04
glucose	0.04	0.01
bmi	0.13	0.02
diabetes_pedigree_func	0.79	7.2
age	0.01	0.005

Źródło: opracowanie własne

Jak widać, największa wariancja dotyczy zmiennej diabete_pedigree_func. Wynika to z kilka faktów: po pierwsze, cukrzyca jest w miarę nowoczesną chorobą i ludzi po prostu wcześniej na nią nie chorowały (jest to związane z rosnącym poziomem życia). Po drugie, z doświadczeń empirycznych, duże wartości tej funkcji oznaczają raczej skłonność do zachorowania na cukrzycę niż prawdopodobieństwo. Także dla zmiennej age oraz pregnancies zostały podwyższone wariancje, ponieważ oryginalna wariancja wydawała się zbyt optymistyczna (<0.01). Pozostałe oczekiwania pozostały na tym samym poziomie, co zakładano w podobnych badaniach.

Rozdział IV: Metoda estymacji a posteriori. Diagnostyka zbieżności

Jako metoda estymacji wybrana została metoda Monte Carlo Hamiltona bez rozrzedzania, o 8 łańcuchach i 3500 iteracji w każdym. Warm-up został ustawiony na poziomie 50% obserwacji.

W wyniku estymacji otrzymano oszacowania rozkładów a posteriori. Rysunek 9 pokazuje jak wyglądały podstawowe statystyki dotyczące estymacji parametrów (średnie, odchylenia, przedziały ufności). Wartości $\hat{R}$ bliskie do 1 wskazują na to, że łańcuch osiągnął zbieżność.

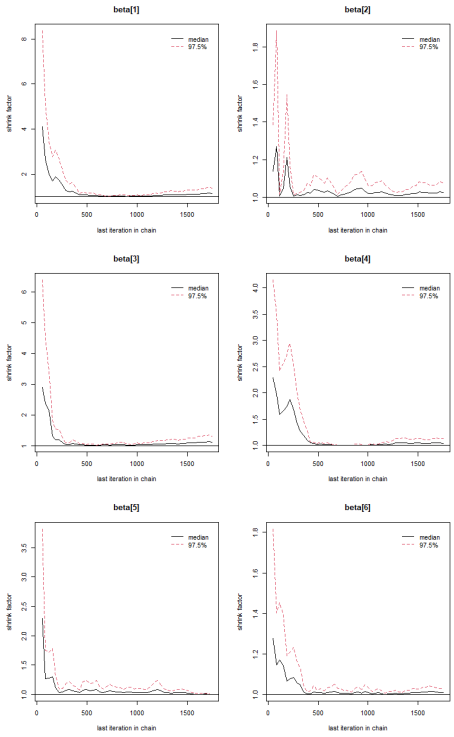
Rysunek 9. Statystyki dotyczące rozkładów a posteriori

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
beta[1]	-9.840000e+00	0.05	0.79	-1.146000e+01	-1.036000e+01	-9.860000e+00	-9.300000e+00	-8.330000e+00	278	1.05
beta[2]	2.300000e-01	0.00	0.03	1.800000e-01	2.100000e-01	2.300000e-01	2.500000e-01	2.900000e-01	1231	1.01
beta[3]	4.000000e-02	0.00	0.00	3.000000e-02	3.000000e-02	4.000000e-02	4.000000e-02	4.000000e-02	444	1.03
beta[4]	1.000000e-01	0.00	0.01	8.000000e-02	9.000000e-02	1.000000e-01	1.100000e-01	1.300000e-01	608	1.02
beta[5]	1.230000e+00	0.01	0.37	5.300000e-01	9.800000e-01	1.220000e+00	1.470000e+00	1.980000e+00	857	1.01
beta[6]	1.000000e-02	0.00	0.00	0.000000e+00	1.000000e-02	1.000000e-02	1.000000e-02	2.000000e-02	1530	1.00
sigma	8.901915e+307	NaN	Inf	3.769049e+306	4.495598e+307	8.761723e+307	1.339332e+308	1.749247e+308	NaN	NaN
lp__	4.642900e+02	0.06	2.03	4.595100e+02	4.631600e+02	4.646300e+02	4.657700e+02	4.672800e+02	1332	1.01

Źródło: opracowanie własne

Z kolei zestaw wykresów 4.2 potwierdza wniosek o zbieżności modelu. Statystyka Gelmana dla 95%-go przedziału ufności nie przekroczyła wartość 1.2. Co więcej, dla 5 z 6 parametrów zbieżność osiągnięta już po 500 iteracjach. Trochę skokowo wygląda zbieżność (shrink factor) parametru przy zmiennej pregnancies, ale i ona z czasem zanika.

Rysunek 10. Statystyka Gelmana



	Point est.	Upper C.I.
beta[1]	1.08	1.17
beta[2]	1.01	1.02
beta[3]	1.06	1.13
beta[4]	1.03	1.06
beta[5]	1.01	1.01
beta[6]	1.01	1.01

Multivariate psrf

1.09

Źródło: opracowanie własne

Warto wyjaśnić niektóre fakty wynikające z wykresów testów zbieżności z wykorzystaniem kryterium Heidelberga-Welcha i testu stacjonarności Cramera-von-Misesa. Wskazują oni na to, że dla większości parametrów zbieżność została osiągnięta. Parametry β_1 oraz β_3 potencjalnie mogą mieć problemy związane ze zbieżnością – zaobserwowane to dla łańcuchu 2. Jednak było to zauważono tylko dla 1 jednego łańcucha i może być związane z nieodpowiednim wyborem wartości startowych.

Rysunek 11. Testy zbieżności

1

	Stationarity test	start iteration	p-value
beta[1]	passed	351	0.282
beta[2]	passed	1	0.628
beta[3]	passed	351	0.128
beta[4]	passed	351	0.576
beta[5]	passed	1	0.936
beta[6]	passed	1	0.348

2

	Stationarity test	start iteration	p-value
beta[1]	failed	NA	0.03061
beta[2]	passed	1	0.20398
beta[3]	failed	NA	0.00282
beta[4]	passed	701	0.08931
beta[5]	passed	1	0.35497
beta[6]	passed	1	0.11322

	Stationarity test	start iteration	p-value		Stationarity test	start iteration	p-value
beta[1]	passed	1	0.886	beta[1]	passed	1	0.607
beta[2]	passed	1	0.420	beta[2]	passed	1	0.133
beta[3]	passed	1	0.336	beta[3]	passed	1	0.342
beta[4]	passed	1	0.973	beta[4]	passed	1	0.336
beta[5]	passed	1	0.849	beta[5]	passed	1	0.392
beta[6]	passed	1	0.675	beta[6]	passed	1	0.397

5

	Stationarity test	start iteration	p-value
beta[1]	passed	176	0.6288
beta[2]	passed	1	0.4406
beta[3]	passed	176	0.5964
beta[4]	passed	1	0.0562
beta[5]	passed	1	0.6523
beta[6]	passed	1	0.7640

6

	Stationarity test	start iteration	p-value
beta[1]	passed	1	0.8053
beta[2]	passed	1	0.0767
beta[3]	passed	1	0.9192
beta[4]	passed	1	0.5142
beta[5]	passed	1	0.6866
beta[6]	passed	1	0.3873

7

	Stationarity test	start iteration	p-value
beta[1]	passed	1	0.1854
beta[2]	passed	1	0.1922
beta[3]	passed	1	0.0884
beta[4]	passed	1	0.2898
beta[5]	passed	1	0.2088
beta[6]	passed	1	0.3217

8

	Stationarity test	start iteration	p-value
beta[1]	passed	1	0.302
beta[2]	passed	1	0.398
beta[3]	passed	1	0.061
beta[4]	passed	1	0.405
beta[5]	passed	1	0.275
beta[6]	passed	1	0.253

Źródło: opracowanie własne

Ostatecznie sprawdzony został dobór liczby iteracji. Przy przyjętym poziomie istotności $\alpha = 0.95$ i precyzji szacunku wynoszącej 0.01, kryterium Raftery'ego sugeruje liczbę iteracji (długość łańcucha) równą co najmniej 9604. Niemniej jednak, ze względu na to że zbieżność została osiągnięta, wyniki testu zostały zignorowane.

Rysunek 12. Test Raftery'ego

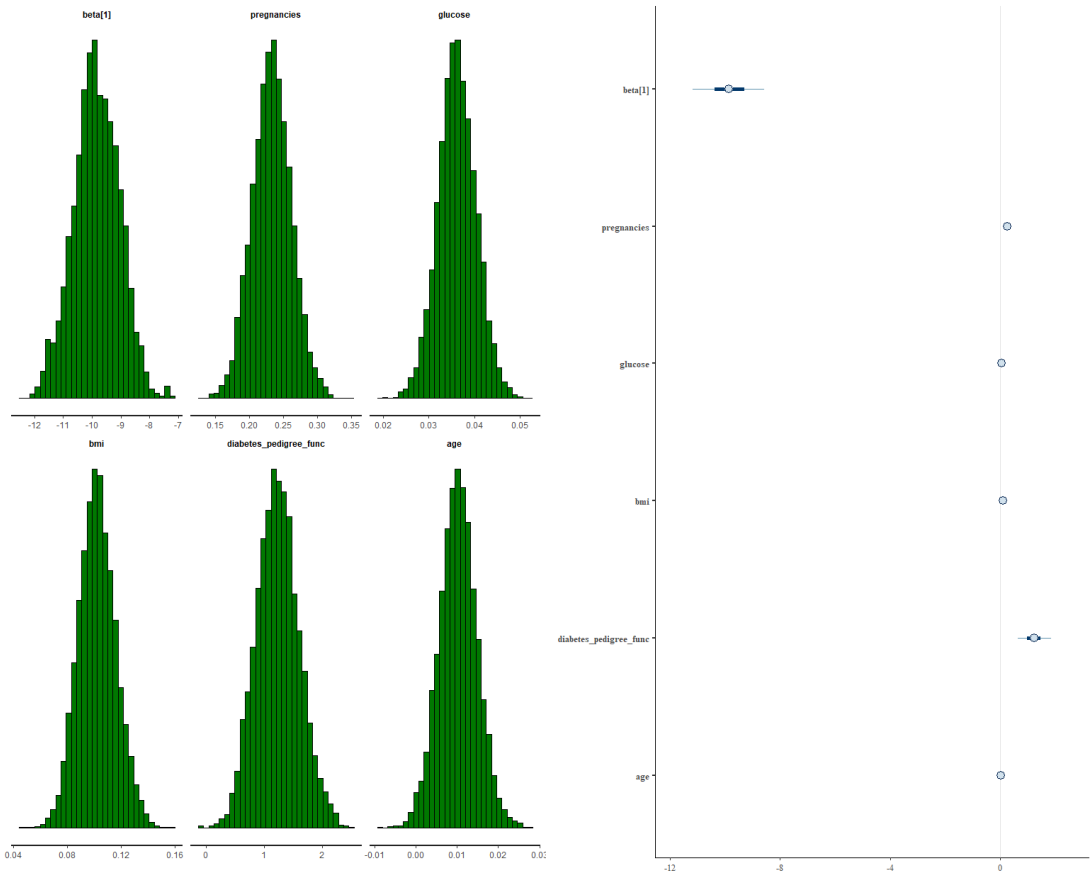
```
Quantile (q) = 0.5
Accuracy (r) = +/- 0.01
Probability (s) = 0.95

You need a sample size of at least 9604 with these values of q, r and s
```

Źródło: opracowanie własne

Rozdział V: Rozkłady a posteriori. HPDI. Czynniki Bayesa

Rysunek 13 przestawia rozkłady a posteriori parametrów modelu. Jak wynika z rysunku, rozkłady są bardzo symetryczne i podobne do rozkładów normalnych. Niemniej jednak, rozkłady dla stałej, zmiennej bmi, diabetes_pedigree_func oraz pregnancies dużo się różnią od założonych rozkładów a priori. W miarę dobrze jest pokrycie z założeniami a priori dla zmiennej age oraz glucose. Wnioski te mogą posłużyć dobrym przykładem dla przyszłych badaczy.

Rysunek 13. Rozkłady a posteriori. HDPI




```
> hpdi.b
                          lower       upper
beta[1]             -11.394871683 -8.27938166
pregnancies           0.178203563  0.28957188
glucose               0.028154224  0.04453605
bmi                   0.074792992  0.12874078
diabetes_pedigree_func 0.526104866  1.97040486
age                   0.001102365  0.01925408
```

Źródło: opracowanie własne

Z kolei Przedziały HPDI sugerują, że wszystkie zmienne istotnie objaśniają czy człowiek jest chory na cukrzycę. Jedyną możliwie nieistotną zmienną jest zmienna age (w przypadku

lewego skrajnego przedziału współczynnik wynosi mniej niż 0.01). Pozostałe zmienne należy uznać za statystycznie istotne niezależnie od sensownie przyjętego poziomu ufności.

Podsumowanie wyników oszacowań są przedstawione w poniższej tabeli:

Tabela 2. Rozkłady a posteriori. HDPI

Zmienna	Rozkład a priori	$\bar{E}$	Istotność HDPI	Czynnik Bayesa
Wyraz wolny	N(-20.76, 46.63)	-9.84	wysoka	-
pregnancies	N(0.33, 0.04)	0.23	średnia	0.01
glucose	N(0.04, 0.01)	0.04	średnia	>6e15
bmi	N(0.13, 0.02)	0.1	średnia	13.03
dpf	N(0.79, 7.2)	1.23	wysoka	0.01
age	N(0.01, 0.005)	0.05	niska	<0.01

Źródło: opracowanie własne

Wyniki oszacowań okazały się kontrowersyjnymi dla większości zmiennych. Wyraz wolny miał znacznie mniejszą średnią a posteriori, niż było założono na początku analizy. Zmienne pregnancies, glucose i bmi, chociaż i (prawie) spełniły założenia a priori, miały różne istotności ze względu na rozkłady HDPI i czynnik Bayesa. W ten sposób np. zmienna glucose okazała się średie ważną według analizy HDPI, ale najbardziej istotną ze względu na BF. Zmienna diabetes_pedigree_func miała wyższą wartość oczekiwaną niż zakładano a priori. Jednak znów wyniki HDPI i BF różniły się. Jedyną zmienną wnioski na temat której zgadzają się dla obu narzędzi interpretacyjnych jest zmienna age.

Podsumowując, w bieżącej prace udało się osiągnąć zarówno zbieżności, jak i większości poczynionych założeń. Zbudowany model sugeruje, że zmienna diabetes_pedigree_func, czyli zmienna odpowiadająca za genealogię pacjenta, jest najistotniejszym czynnikiem – im większa jest bliskość czasowa i krewna do ostatniego przypadku choroby na cukrzycę wśród rodziny człowieka, tym większe jest prawdopodobieństwo, że osoba ta też będzie chora na cukrzycę. Niemniej jednak, wyniki te są dość sporne w porównaniu do tradycyjnych czynników których przyjęto liczyć najistotniejszymi (blood_pressure, glucose). Dlatego należy przeprowadzić dodatkowe estymacje z wykorzystaniem większych i kompletniejszych zbiorów danych zanim zacząć wdrażać wyciągnięte wnioski w praktyce.

Bibliografia

1. Kapat P., Wang K, *Classification Using Bayesian Logistic Regression: Diabetes in Pima Indian Women Example*, https://www.asc.ohio-state.edu/goel.1/STAT825/PROJECTS/KapatWang_Team4Report.pdf (dostęp 07.02.2023).
2. Dritsas E., Trigka M., *Data-Driven Machine-Learning Methods for Diabetes Risk Prediction*, University of Patras, Patras 2022.
3. Hassan M., *A fully bayesian logistic regression model for classification of zada diabetes dataset*, University of Zakho, Iraq 2020.
4. https://diabetesjournals.org/care/article/41/5/963/36522/Global-Economic-Burden-of-Diabetes-in-Adults, (dostęp 07.02.2023).
5. https://www.kaggle.com/datasets/mathchi/diabetes-data-set?datasetId=818300&sortBy=voteCount , (dostęp 07.02.2023).

Spis rysunków

Spis tabeli