Analiza danych statystycznych dotyczacych zachorowan na cukrzyce

Adam Dyda
AGH, Wydzial Informatyki Elektroniki i Telekomunikacji
Rachunek prawdopodobienstwa i statystyka 2020/2021

Krakow, 19 stycznia 2021

Adam Dyda

Spis treści

1 Opis danych

Analizowane dane znajduja sie na stronie `https://www.kaggle.com/uciml/pima-indians-diabetes-database`. Pochodza one z instytutu *National Institute of Diabetes and Digestive and Kidney Diseases*. Dane dotycza zachorowan na cukrzyce w grupie kobiet Indian Pima *(ang. Pima Indian)* powyzej 21 roku zycia. Zawieraja one informacje odnoszace sie do kilku medycznych pomiarow i cech osob poddawanych analizie. Kolumny zbioru danych ktory bede analizowal:

- Pregnancies - liczba ciazy
- Glucose - poziom glukozy we krwi mierzyony podczas dwugodzinnego doustnego testu obciazenia glukoza
- BloodPreasure - rozkurczowe cisnienie krwi (mm Hg)
- SkinThickness - grubosc faldu skornego na tricepsie (mm)
- Insulin - poziom insuliny we krwi
- BMI - wskaznik masy ciala
- DiabetesPedigreeFunction (DPF) - funkcja okreslajaca szanse na na zachorowanie na podstawie historii przodkow (ich zachorowan na cukrzyce) oraz relacji genetycznych z przodkami
- Age - wiek
- Outcome - cecha wskazujace czy dana osoba ma cukrzyce (0 lub 1)

2 Czyszczenie danych

Pierwszym etapem naszej analizy bedzie sprawdzenie czy dane sa kompletne i nie zawieraja zadnych brakujacych wartosci, przyjrzyjmy sie ich podsumowaniu.

```
  Pregnancies        Glucose      BloodPressure    SkinThickness
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
    Insulin           BMI             DPF              Age
 Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00
 1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00
 Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
 Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
 3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
 Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
```

```
    Outcome
 Min.   :0.000
 1st Qu.:0.000
 Median :0.000
 Mean   :0.349
 3rd Qu.:1.000
 Max.   :1.000
```

Jak widac dane nie zawieraja brakujacych wartosci (mozna to zaobserwowac poprzez brak wpisu "NA's" w ktorejkolwiek z kolumn). Mozemy jednak zauwazyc ze niektore z kolumn posiadaja 0 jako wartosc minimalna. Zastanowmy sie teraz czy dla wszystkich kolumn wartosc 0 moze byc rzeczywiscie zmierzona i jest mozliwa do uzyskania. Poziom glukozy, cisnienie krwi, poziom insuliny, grubosc skory oraz BMI nie moga przyjmowac wartosci 0. Sprawdzmy teraz ile jest wartosci 0 w calym zbiorze danych.

```
> colSums(df == 0)
```

```
  Pregnancies       Glucose BloodPressure SkinThickness       Insulin
          111             5            35           227           374
          BMI           DPF           Age       Outcome
           11             0             0           500
```

Wywnioskowac z tego mozemy ze w zbiorze brak danych jest oznaczany wartoscia 0 (taki wniosek przyjmujemy dla kolumn wczesniej wymienionych dla ktorych ta wartosc "nie ma sensu" w pozostalych kolumnach uznajemy wartosc 0 jako wartosc poprawna)

Brakujace dane nalezy odrzucic lub uzupelnic, odrzucenie wierszy z brakujacymi danymi w tym wypadku wiaze sie z utrata bardzo duzej ilosci informacji poniewaz dla kolumny *Insulin* sa to az 374 wartosci, rozwazyc mozna takze odrzucenie calej kolumny jednak informacja o poziomie insuliny jest dla nas zbyt istotna przy badaniu zachorowan na cukrzyce.

Zdecydowalem sie na uzuplenienie brakujacych danych odpowiednimi wartosciami w zaleznosci od kolumny. W przypadku kolumn z dostatecznie mala skosnoscia rozkladu brakujace dane bede uzupelnial srednia arytmetyczna z pozostalych poprawnych danych w tej kolumnie, w przypadku rozkladu z duza skosnoscia uzyje mediany do uzupelnienia danych. Sprawdzmy skosnosc wszystkich kolumn w zbiorze:

```
> skewness(df)
```

```
  Pregnancies       Glucose BloodPressure SkinThickness       Insulin
    0.8999119     0.1734140    -1.8400052     0.1091588     2.2678105
          BMI           DPF           Age       Outcome
   -0.4281433     1.9161592     1.1273893     0.6337757
```

Za skosnosc dostateczna do uzupelnienia danych srednia artytmetyczna przyjmuje wartosc wieksza od -0.5 oraz mniejsza od 0.5. Oczywiscie uzupelnienie brakujach danych srednia lub mediana wprowadza przeklamania z ktorych nalezy zdawac sobie sprawe, dlatego tez w dalszej czesci analizy jezeli bedzie to potrzebne bede uzywal danych oryginalnych z odrzuconymi danymi brakujacymi (dotyczy to glownie kolumny *Insulin*). Taki zabieg zostanie odpowiednio zasygnalizowany wczesniej.

3 Analiza danych

Zajmiemy sie teraz analiza rozkladu danych w naszym zbiorze, przyjrzymy sie histogramom oraz dokonamy estymacji parametrow naszych rozkladow za pomoca stymatora MLE *(ang. Maximum likelihood estimator)*. Nastepnie przejdziemy do analizy boxplotow oraz powiemy o wartosciach odstajacych w naszych danych.

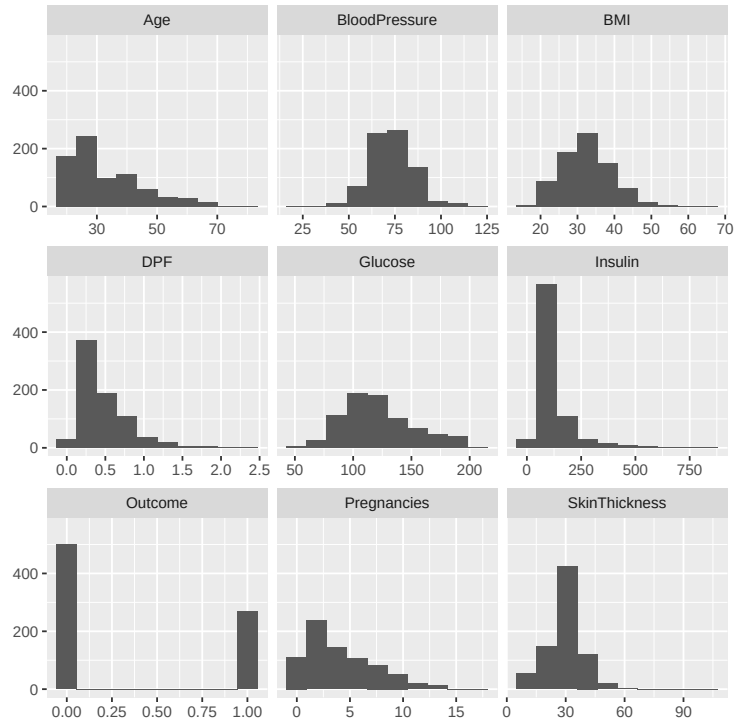
3.1 Analiza rozkladu danych

Analize zaczniemy od ponownego przeanalizowania rezultatu polecenia *summary* tym razem juz dla uzupelnionych danych:

```
  Pregnancies        Glucose       BloodPressure    SkinThickness
Min.   : 0.000   Min.   : 44.00   Min.   : 24.00   Min.   : 7.00
1st Qu.: 1.000   1st Qu.: 99.75   1st Qu.: 64.00   1st Qu.:25.00
Median : 3.000   Median :117.00   Median : 72.20   Median :29.00
Mean   : 3.845   Mean   :121.66   Mean   : 72.41   Mean   :29.11
3rd Qu.: 6.000   3rd Qu.:140.25   3rd Qu.: 80.00   3rd Qu.:32.00
Max.   :17.000   Max.   :199.00   Max.   :122.00   Max.   :99.00
    Insulin           BMI             DPF              Age
Min.   : 14.0   Min.   :18.20   Min.   :0.0780   Min.   :21.00
1st Qu.:121.5   1st Qu.:27.50   1st Qu.:0.2437   1st Qu.:24.00
Median :125.0   Median :32.30   Median :0.3725   Median :29.00
Mean   :140.7   Mean   :32.46   Mean   :0.4719   Mean   :33.24
3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
    Outcome
Min.   :0.000
1st Qu.:0.000
Median :0.000
Mean   :0.349
3rd Qu.:1.000
Max.   :1.000
```

Na podstawie tych informacji mozna zaobserwowac ze w przypadku kazdej z cech rozklad jest prawostronnie skosny (srednia arytmetyczna jest wieksza od mediany). W przypadku kolumn *Insulin, Pregnancies, DPF, SkinThickness* widzimy duza roznice pomiedzy trzecim kwartylem a wartoscia maksymalna co

moze oznaczac jedna lub potencjalnie wiele wartosci odstajacych.
Przyjrzyjmy sie teraz histogramom naszych danych:



Na powyzszych histogramach widzimy ze kolumny *BloodPressure, Glucose* oraz *BMI* posiadaja rozklad ktory przypomina rozklad normalny. A wiec wlasnie tymi cechami zajmiemy sie w kolejnej sekcji podczas estymacji parametrow rozkladu.
Warto takze zwrocic uwage na rozklad danych w kolumnie *Outcome* i zauwazyc ze nie sa one rownomiernie rozlozone. Sprawdzmy dokladna liczbe osob chorych i zdrowych.

```
> table(df$Outcome)

  0   1
500 268
```

Jak widac liczba osob zdrowych jest prawie dwa razy wieksza od liczby osob chorych, co moze miec znaczenie w dalszej analizie.

3.2 Estymacja parametrow rozkladu

Teraz zajmiemy sie estymacja parametrow μ oraz σ^2 rozkladow trzech cech ktore wybralismy w poprzednim punkcie tj. *BloodPressure, Glucose* oraz *BMI*. Estymacji bedziemy dokonywac za pomoca estymatora najwiekszej wiarygodnosci

MLE. Na poczatku zdefiniujmy sobie funkcje wiarygodnosci, w naszym wypadku jest to *Negative Likelihood function*, co oznacza ze bedziemy minimalizowac te funkcje.

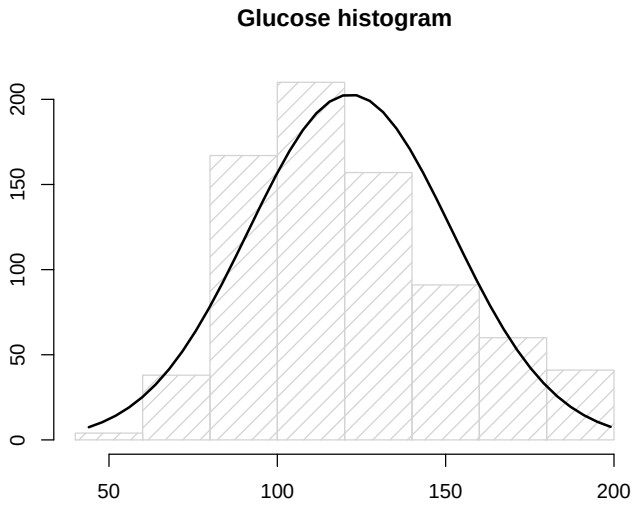
```
> NLLGlucose <- function(theta0,theta1) {
+   -sum ( -0.5* log(theta1*2*pi) - 0.5*( df$Glucose- theta0)^2/theta1 )
+ }
```

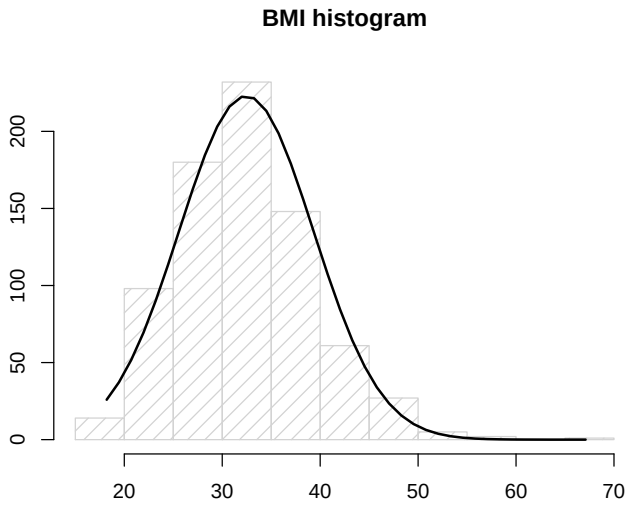
Analogicznie definiujemy funkcje dla kolumn *BloodPressure* oraz *BMI*. Nastepnie minimalizujemy funkcje wiarygodnosci aby otrzymac odpowiednie parametry, uzywamy do tego funkcji *mle* z pakietu *stats4*

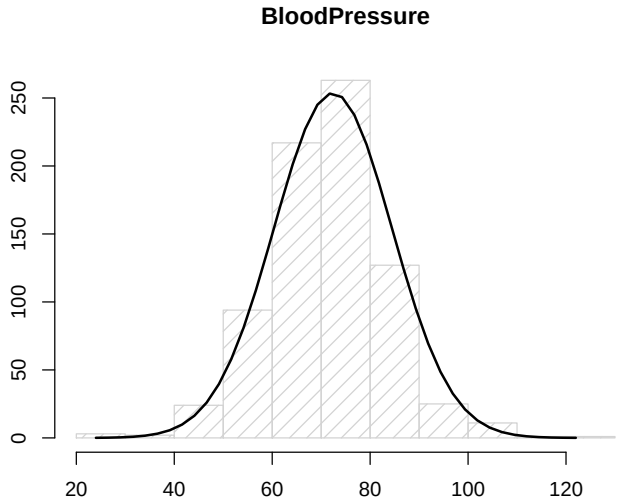
theta0 oznacza parametr μ natomiast theta1 oznacza σ^2

```
> Glucoseest <- stats4::mle(minuslog=NLLGlucose, start=list(theta0=100,theta1=900))
> coef(Glucoseest)

  theta0   theta1
121.6624 913.1354

> BMIest <- stats4::mle(minuslog=NLLBMI, start=list(theta0=30,theta1=60))
> coef(BMIest)

  theta0   theta1
32.45521 47.20657

> BPest <- stats4::mle(minuslog=NLLBP, start=list(theta0=70,theta1=375))
> coef(BPest)

   theta0    theta1
 72.40497 145.92579
```

Na podstawie wyestymowanych informacji narysujmy wykresy rozkladow normalnych z otrzymanymi parametrami na histogramach kazdej z cech.


Glucose histogram
0
50
100
150
200
50
100
150
200


BMI histogram
0
50
100
150
200
20
30
40
50
60
70



Wyestymowane rozklady wygladaja na bliskie prawdy tj. takie ktorych parametry bliskie sa rzeczywistym parametrom rozkladu. Porownajmy wiec wartosci rzeczywiste z wyestymowanymi.

Rzeczywiste wartosci parametrow dla kolumny *Glucose*

```
> mean(df$Glucose)

[1] 121.6562

> var(df$Glucose)

[1] 926.4892
```

Estymowane wartosci parametrow dla kolumny *Glucose*
theta0 oznacza parametr μ natomiast theta1 oznacza σ^2

```
  theta0   theta1
121.6624 913.1354
```

Rzeczywiste wartosci parametrow dla kolumny *BMI*

```
> mean(df$BMI)

[1] 32.45521

> var(df$BMI)

[1] 47.26806
```

Estymowane wartosci parametrow dla kolumny *BMI*
theta0 oznacza parametr μ natomiast theta1 oznacza σ^2

```
  theta0   theta1
32.45521 47.20657
```

Rzeczywiste wartosci parametrow dla kolumny *BloodPressure*

```
> mean(df$BloodPressure)

[1] 72.40518

> var(df$BloodPressure)

[1] 146.3216
```

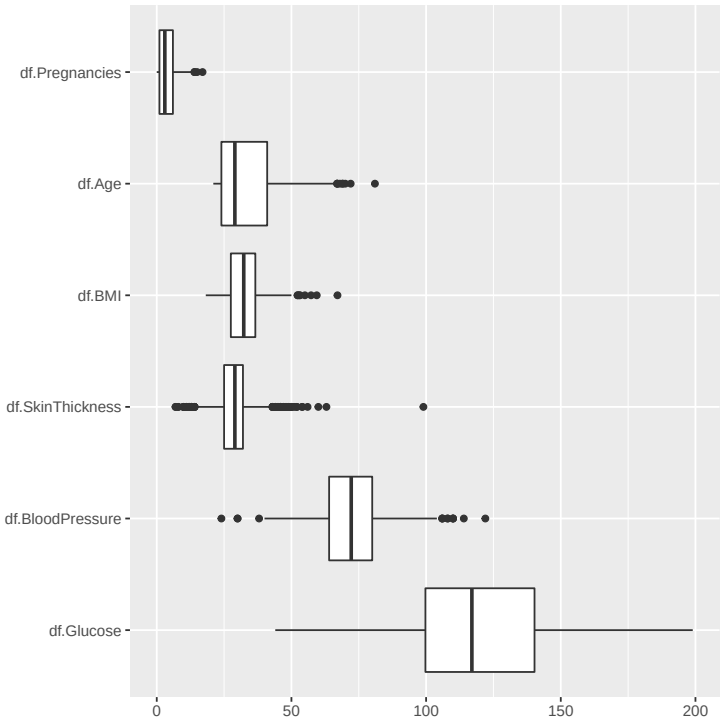
Estymowane wartosci parametrow dla kolumny *BloodPressure*
theta0 oznacza parametr μ natomiast theta1 oznacza σ^2

```
   theta0    theta1
 72.40497 145.92579
```

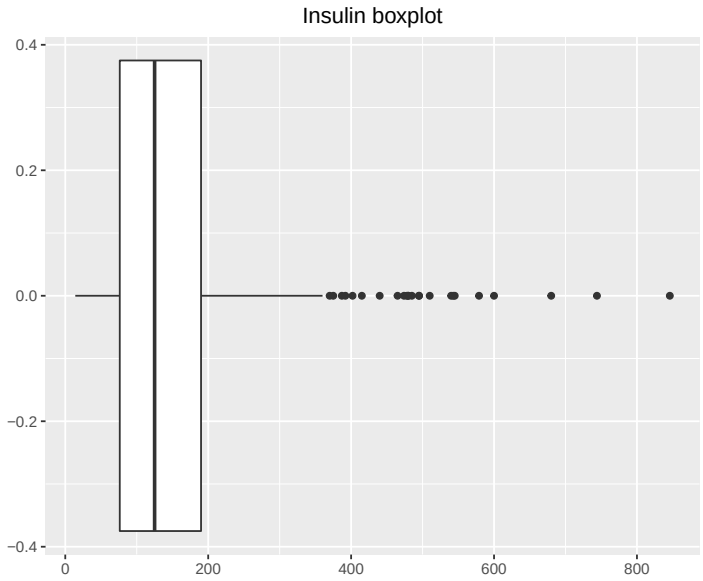
Estymowane wartosci sa bardzo bliskie prawdy, mozna wiec stwierdzic ze estymacja za pomoca metody najwiekszej wiarygodnosci okazala sie wlasciwa w tym przypadku.

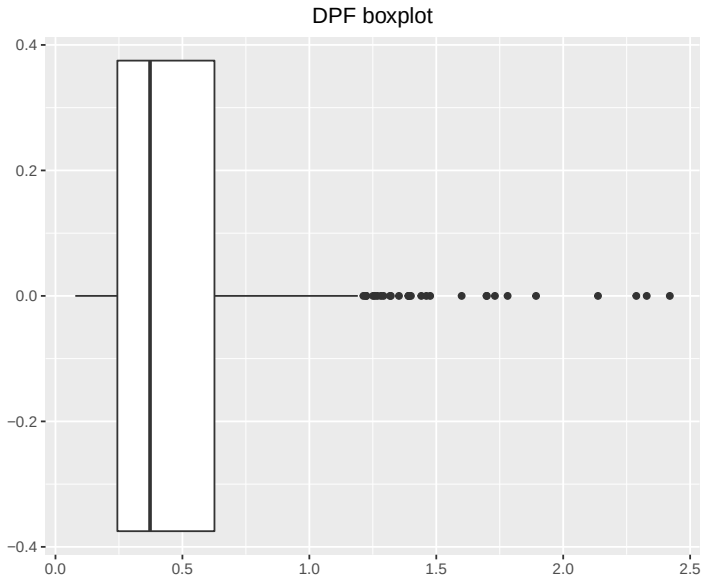
3.3 Analiza wykresow pudelkowych

Dalsza czescia eksploracji naszych danych bedzie narysowanie oraz analiza wykresow pudelkowych kazdej z cech. Wykresy rysuje w trzech oddzielnych grupach poniewaz wartosci sa z roznych rzedow wielkosci i narysowane na jednym wykresie bylyby nieczytelne.



Podczas rysowania wykresu pudelkowego kolumny *Insulin* zostaly uzyte dane z usunietymi wierszami ktorych wartosc w oryginalnym zbiorze wynosila 0.





Na powyzszczych wykresach mozemy zauwazyc znaczaca ilosc wartosci odstajacych w kolumnach *Insulin, DPF* oraz *SkinThickness*, co moze miec wplyw na zaklamania podczas uzywania regresji liniowej w dalszej czesci analizy. Wynika to z tego ze model regresyjny jest bardzo czuly na wartosci odstajace.

Sprawdzmy dokladna liczbe wartosci odstajacych w tych kolumnach. Za wartosci odstajace przyjmujemy wartosci mniejsze od $Q1 - 1.5IQR$ oraz wieksze od $Q3 + 1.5IQR$ gdzie $Q1$,$Q3$ to odpowiednio wartosci pierwszego i trzeciego kwartyla a IQR to rozstep cwiartkowy czyli roznica miedzy trzecim i pierwszym kwartylem.

```
> length(boxplot(cleandf$Insulin)$out)

[1] 24

> length(boxplot(df$DPF)$out)

[1] 29

> length(boxplot(df$SkinThickness)$out)

[1] 87
```

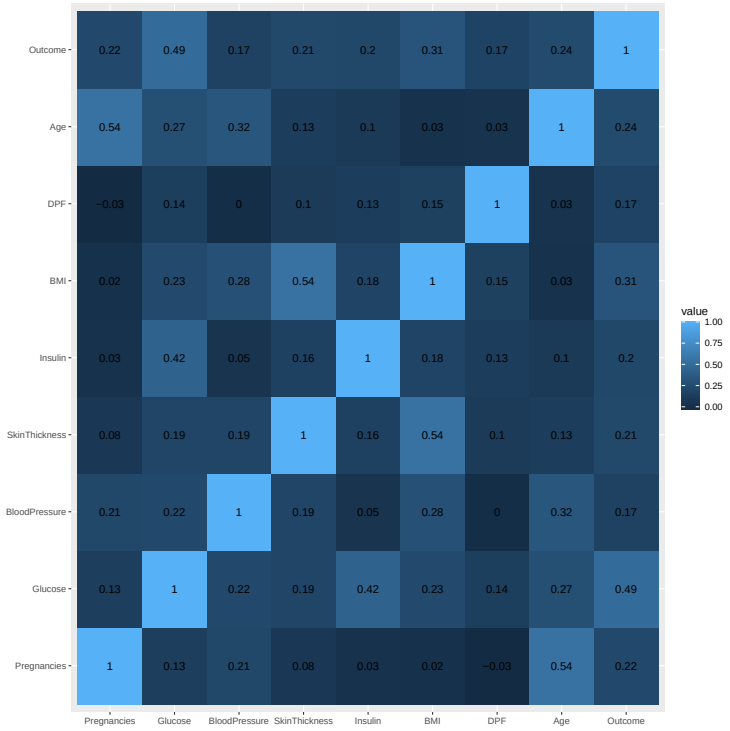
Warto zauwazyc ze wartosc dla kolumny *SkinThickness* jest zawyzona poniewaz uzywamy tutaj danych z uzupelnionymi brakujacymi wartosciami

4 Analiza korelacji oraz zaleznosci danych

W tym kroku zajmiemy sie analiza zaleznosci pomiedzy danymi za pomoca wspolczynnika korelacji Pearsona. W szczegolnosci zajmiemy sie badaniem korelacji pomiedzy kolumna *Outcome* a pozostalymi cechami w celu ustalenia ktore z cech maja najwieksza korelacje z zachorowaniem na cukrzyce.

4.1 Analiza korelacji za pomoca wspolczynnika Pearsona

W celu zwizualizowania korelacji miedzy kolumnami wykorzystam wykres typu heatmap ktory w jednym miejscu przedstawia zaleznosci pomiedzy wszystkimi kolumnami.



Na pierwszy rzut oka mozemy zobaczyc np. wysoka korelacje wieku z liczba ciaz lub wieku z cisnieniem krwi. Sa to oczywiste zaleznosci ktorymi nie bedziemy sie glebiej zajmowac jednakze warto na nie zwrocic uwage. Wartosci ktore nas interesuja jest to korelacja zachorowan na cukrzyce z reszta kolumn, w tym wypadku dwie najwieksze wartosci czyli korelacje z poziomem glukozy we krwi oraz wskaznikiem BMI, interesuje nas takze korelacja zachorwan na cukrzyce z poziomem insuliny we krwi poniewaz jest to bardzo czesto badana cecha wsrod chorych na te chorobe.

4.2 Regresja liniowa oraz analiza korelacji

W tym punkcie uzyjemy modelu regresji liniowej do wizualizacji zaleznosci pomiedzy wybranymi cechami. Pierwszym wykresem bedzie wykres zaleznosci poziomu glukozy do zachorowan na cukrzyce do ktorego dopasowujemy prosta za pomoca regresji liniowej.

Jak widzimy korelacja jest poztywna co oznacza ze im wiekszy poziom glukozy we krwii tym wieksza szansa na to ze osoba jest chora na cukrzyce. Jest to wynik ktorego sie spodziewalismy poniewaz czesto wysoki poziom cukru jest wlasnie indykatorem cukrzycy typu 1 lub 2.

Nastepnie narysujmy za pomoca modelu regresji liniowej wykres zaleznosci wskaznika BMI do zachorowan na cukrzyce

W tym wypadku korelacja takze jest pozytywna co oznacza ze im wiekszy wskaznik BMI tym wieksza szansa zachorowania na cukrzyce. Jest to wniosek ktory potwierdza ogolne przekonanie ze osoby otyle czesciej choruja na cukrzyce. Wykresem kolejnej zaleznosci bedzie zaleznosc poziomu insuliny we krwi do zachorowan na cukrzyce. W tym wypadku uzywam zbioru danych z usunietymi wierszami ktorych wartosc poziomu insuliny w oryginalnym zbiorze wynosila 0.

Zaleznosc w tym wypadku takze jest pozytywna co jest zaskakujaca obserwacja, poniewaz spodziewamy sie ze osoby chore na cukrzyce maja niski poziom insuliny. W naszym wypadku mozna wyciagnac dalsze wnioski z tej obserwacji, biorac pod uwage ze osoby chore na cukrzyce typu 1 maja niski poziom insuliny natomiast osoby chore na cukrzyce typu 2 moga miec ten poziom insuliny podwyzszony (co nie jest w tym wypadku regula) m.in. dlatego ze hiperinsulinemia moze prowadzic do rozwoju cukrzycy typu 2. Mozemy wywnioskowac ze dane osob chorych w naszym zbiorze pochodza od osob w wiekszosci (lub wszystkich) chorych na cukrzyce typu 2.

Warto zaznaczyc ze powyzsze dopasowania lini za pomoca regresji liniowej sa niedokladne z kilku powodow. Regresja liniowa sprawdza sie najlepiej w wypadku w ktorym cechy pochodza z rozkladu normalnego, w naszym zbiorze kolumna *Outcome* przyjmuje tylko dwie wartosci, wiec zaniza to poziom skutecznosci naszego modelu. Kolejnym powodem jest juz wczesniej wspomniany problem z wartosciami odstajacymi na ktore model regresji liniowej jest bardzo czuly, jednym z rozwiazan tego problemu mogloby byc uzycie regularyzowanej regresji liniowej, czyli regresji z uzyciem parametru penalizujacego. *(Ridge Regression, Lasso Regression)*

5 Wnioski

Analiza danych zakonczyla sie sukcesem, udalo nam sie sprawnie wyczyscic dane i doprowadzic je do stanu w ktorym sa uzyteczne i zdatne do analizowania. Dalsza analiza pozwolila nam odkryc asymetrycznosci i nierownomiernosci w naszych danych. Kolejny etap czyli estymacja parametrow na pomoca metody

MLE takze dala zamierzony efekt, poniewaz udalo nam sie sprawnie i zadowalajaco dokladnie wyestymowac parametry badanych rozkladow. Przyjrzelismy sie wykresom pudelkowym co pozwolilo okreslic ktore z cech posiadaja wartosci odstajace ktore sa potencjalnym problemem w dalszych etapach analizy. Nastepnie przeszlismy do analizy korelacji miedzy cechami i wyciagania wnioskow na temat zaleznosci miedzy nimi.

Glowna zaleznoscia ktora chcielismy zbadac jest korelacja zachorowan na cukrzyce z cechami osoby potencjalnie chorej. Analizujac nasze wyniki mozna dojsc do wniosku ze badanie zakonczylo sie sukcesem, poniewaz zauwazylismy kilka pozytywny korelacji z zachorowaniami na cukrzyce m.in poziom glukozy we krwi oraz wskaznik BMI co potwierdza powszechna opinie ze poziom glukozy nalezy kontrolowac oraz ze osoby otyle czesciej choruja na cukrzyce. Wyjatkowo ciekawy jest wniosek plynacy z analizy korelacji insuliny z zachorowaniem na cukrzyce. Z tej zaleznosci wywnioskowalismy ze osoby chore w naszym zbiorze danych choruja glownie na cukrzyce typu 2, za ta teza przemawiac moze takze fakt ze dane sa pobrane od zamknietej grupy osob o okreslonej charakterystyce co predysponuje taka grupe do chorowania na wezszy zakres chorob.

Po zakonczeniu analizy danych nalezy zdawac sobie sprawe z niedokladnosci i przeklaman ktore w niej wystepuja oraz ktore wprowadzilismy poprzez uzupelnienie brakujacych danych. Pod uwage nalezy wziac takze to ze zarowno wspolczynnik korelacji Pearsona jak i model regresji liniowej swoja najlepsza efektywnosc uzyskuja dla danych ciaglych o rozkladzie normalnym, jednak nie wszystkie nasze cechy takie wlasciwosci posiadaly. W szczgolnosci kolumna *Outcome* jest cecha jakosciowa (tj. niemierzalna, przyjmujaca wartosci 0 lub 1). Przeklamania wprowadza takze liczba wartosci odstajacych w niektorych kolumnach. Warto sie zastanowic czy w analizie tych danych nie lepszym rozwiazaniem byloby uzycie innych modeli statystycznych, mogloby to pozwilic nam w uzyskaniu dokladniejszych wynikow i uniknac przeklaman i niedokladnosci w niektorych aspektach analizy.