Indiani Pima

EMANUELE DI PIETRO

Il dataset preso in esame fa riferimento ad una popolazione di Indiani, chiamati Indiani Pima, situati a sud-ovest degli Stati Uniti, in particolare nello stato dell'Arizona. È composto da 768 osservazioni, tutte donne di età compresa tra 21 e 81 anni.

Scopo dell'analisi è di vedere eventuali relazioni tra il diabete, l'età e l'IMC (Indice di massa corporea), ossia il rapporto tra Peso (in Kg) e Altezza^2 (in m^2), in particolar modo si vuole vedere se è possibile prevedere l'avere (oppure no) il diabete sulla base delle variabili citate su.

Il dataset contiene 3 variabili ed è stato preso da D.A.S.L.:

https://dasl.datadescription.com/datafile/pima-indians/?_sf_s=pima&_sfm_cases=4+59943

-Diabete: variabile qualitativa dicotomica (0 =Diabete No, 1=Diabete Si)

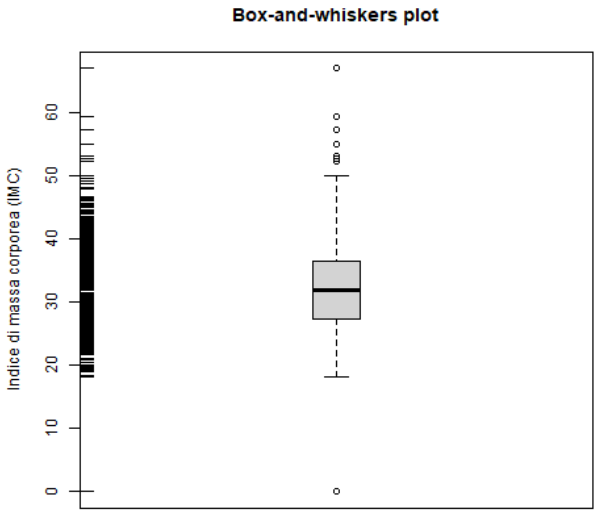
-IMC (Indice massa corporea): Indice di massa corporea

-Età: anni delle donne

Quest'ultima variabile è stata anche suddivisa in classi, per l'esattezza 5: (21,25), (26, 31), (32, 41), (42,50), (50+).

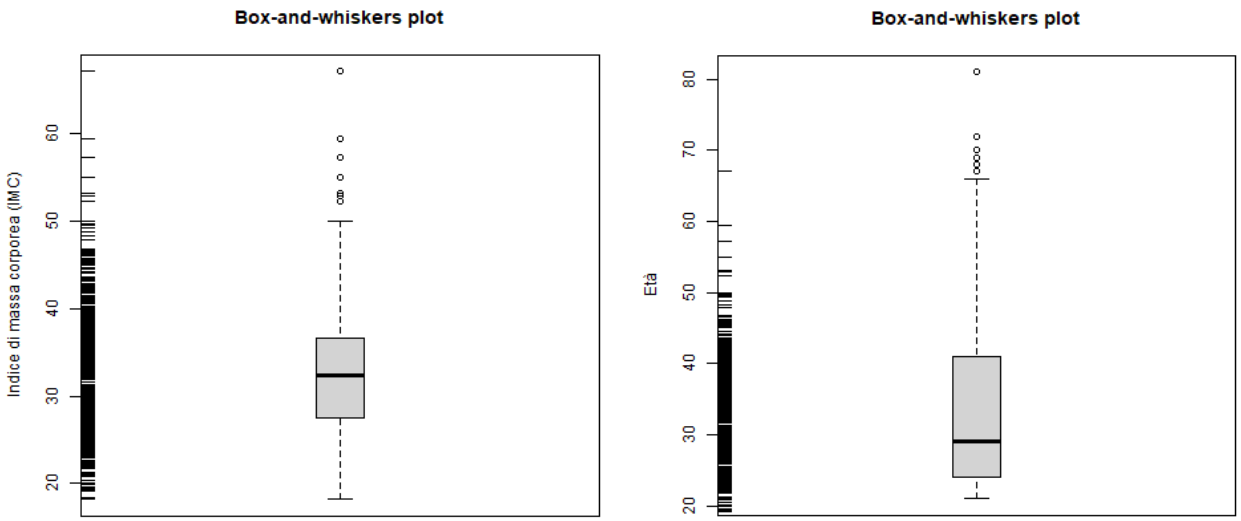
Analisi esplorativa univariata

Per l'analisi esplorativa è conveniente iniziare dalla visualizzazione grafica dei boxplot delle variabili quantitative.



Dal primo boxplot notiamo subito la presenza di valori anomali a ridosso dello 0, probabilmente dovuti ad errori di rilevazione dato che è palesemente impossibile pensare ad un IMC = 0. Quindi, è stato deciso, prima di continuare, di provare a togliere le osservazioni con IMC = 0 per evitare che queste influenzassero l'analisi.

A seguito di ciò la numerosità campionaria è passata da 768 a 757.



Il boxplot adesso non presenta più quei valori anomali pari a 0; analizzando il grafico notiamo il 50% delle osservazioni sono comprese tra 27.50 e 36.60, valori che fanno riflettere, dato che con un IMC ≥ 25.00 si è definiti in “Sovrappeso”. La media e la mediana sono pressoché simili, rispettivamente 32.46 e 32.30, valori ancora più preoccupanti considerando che con un IMC ≥ 35.00 si è in stato di “Obesità di classe II”; infine, notiamo altri valori anomali che superano l’IMC = 50.

Il secondo boxplot, dato che fa riferimento ad una variabile demografica, ci dice solamente com’è composta la popolazione, il 50% delle osservazioni centrali sono comprese tra i 24 e i 41 anni, con valore mio e mediano di 33.28 e 29 anni, una popolazione prevalentemente giovane; notiamo però che ci sono donne che arrivano anche a 70-80 anni.

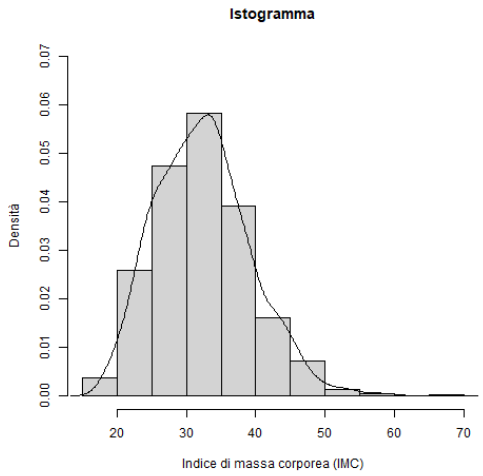
Di seguito è riportata una tabella, presa dal web, per far capire meglio i valori dell'IMC.

Classificazione	BMI (kg/m^2)
Sottopeso	< 18.5
Normopeso	18.5 - 24.9
Sovrappeso	≥ 25.0
pre-obeso	25.0 - 29.9
obeso classe I	30.0 - 34.9
obeso classe II	35.0 - 39.9
obeso classe III	≥ 40.0

Successivamente sono stati stimati alcuni indici di sintesi riguardo la variabile quantitativa IMC, i risultati ottenuto sono i seguenti:

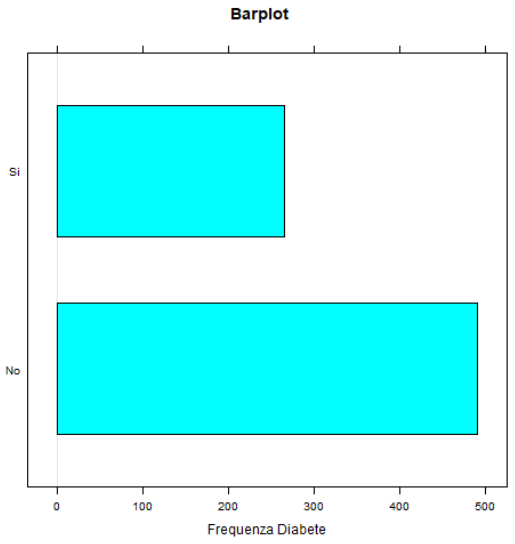
```
> variance(BMI)
[1] 47.89211
> skewness(BMI)
[1] 0.5927921
> kurtosis(BMI)
[1] 3.849771
```

Dalla stima notiamo una piccola asimmetria positiva (a destra), che era possibile notare già dal boxplot, pari a 0.59, mentre vediamo la curtosi è 3.85, forse condizionata dai valori anomali lungo la coda destra. Per capire meglio la forma della funzione di densità è opportuno fare l'istogramma.

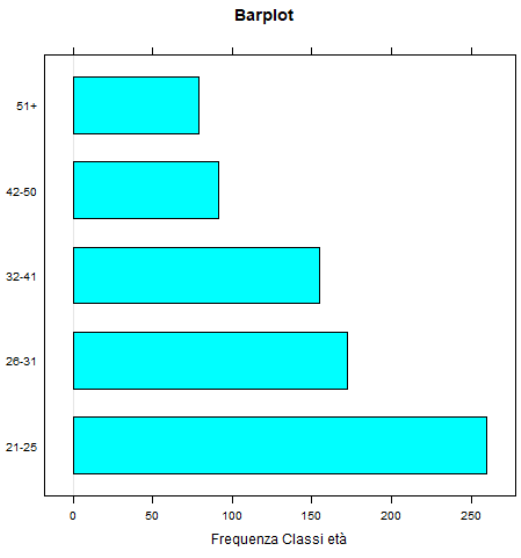


Dall'istogramma e dalla funzione di densità possiamo vedere l'asimmetria positiva, inoltre, se non fosse stato per i valori anomali lungo la coda destra probabilmente tali dati avrebbero seguito una distribuzione normale.

Per quanto riguarda le variabili qualitative invece, è opportuno costruire dei 'barplot':




```
> table(Diabetes)
Diabetes
 No  Si
491 266
```

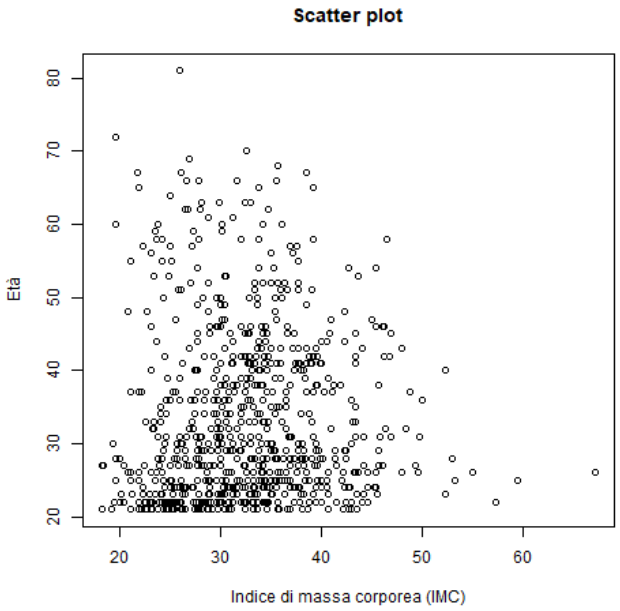



```
> table(Age_Class)
Age_Class
21-25 26-31 32-41 42-50   51+
  260   172   155    91    79
```

Questi grafici ci dicono solamente quante unità sono presenti in ciascun sottogruppo.

Analisi bivariata

Dopo aver fatto un'analisi prendendo le variabili singolarmente bisogna andare a vedere anche come si comportano le variabili prese in coppia:

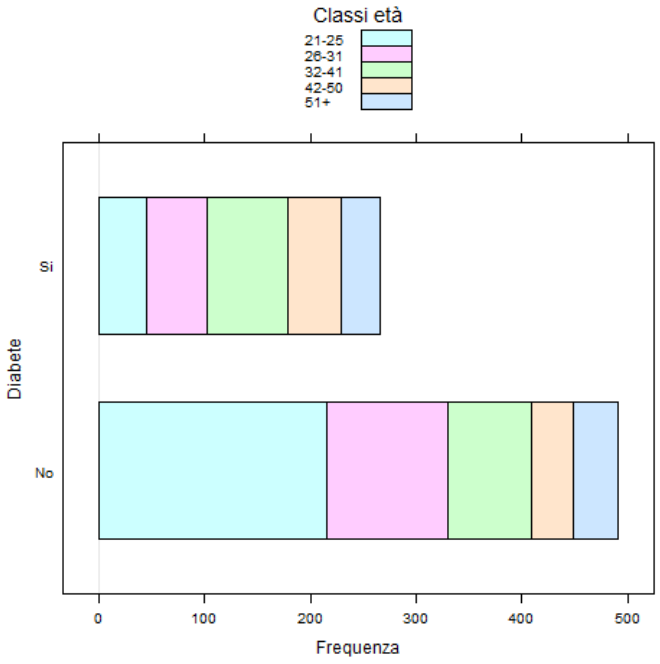


Dallo 'scatter plot' notiamo che non c'è una grande relazione lineare tra l'IMC e l'età, infatti si hanno valori molto elevati dell'IMC in qualsiasi età, anche se i più alti vengono raggiunti da donne molto giovani (≤ 25).

```
> cor(BMI, Age)
[1] 0.02584146
```

La correlazione tra le due variabili è, come si può vedere, 0.026, valore decisamente molto basso.

Per quanto riguarda la coppia di variabili qualitative si adotta un ‘diagramma a nastri condizionato’ accompagnato dalla tabella a doppia entrata delle due variabili.

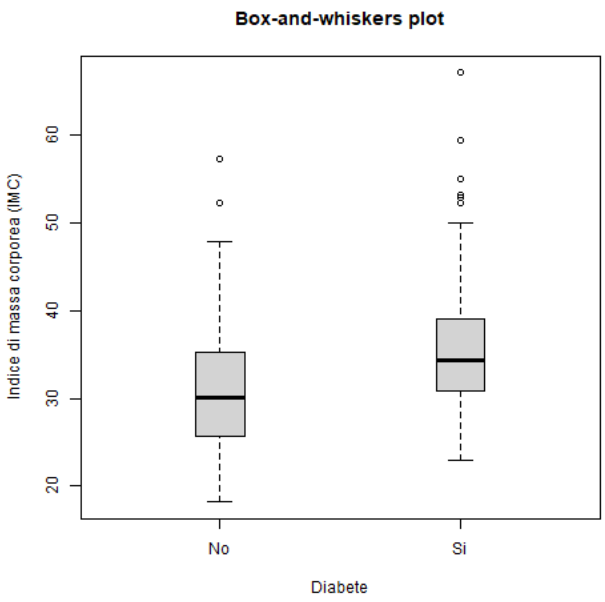


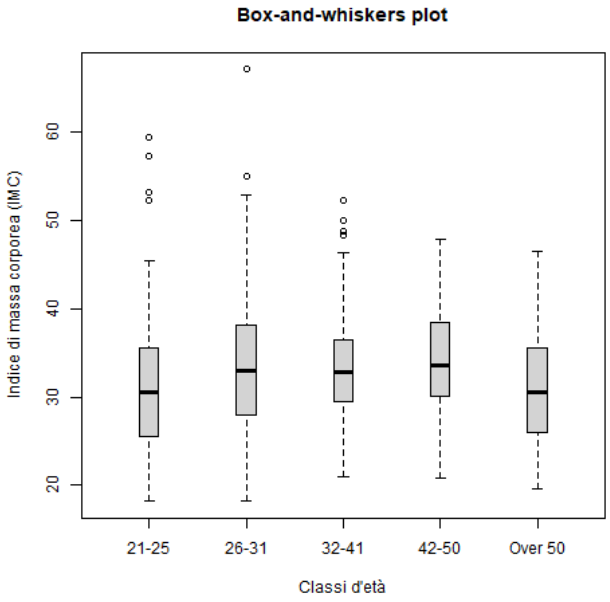

```
> table(Diabetes, Data$Age_Class)

Diabetes 21-25 26-31 32-41 42-50 51+
      No   215   115    79    40  42
      Si    45    57    76    51  37
```

Tale grafico ci informa solo sulla numerosità si ogni variabile condizionata all’altra, in particolar modo possiamo notare che la maggior parte delle donne con diabete fa parte della categoria 32-41 anni, mentre la maggior parte di quelle senza diabete della categoria 21-25 anni; ovviamente tutto ciò è relativo, è conseguenza soprattutto della diversa numerosità campionaria tra le varia classi d’età (quindi più difficili da confrontare). L’unica classe dove le donne con diabete sono maggiori è quella dei 42-50 anni.

Analisi condizionata

Per analisi condizionata si intende analizzare il comportamento di una variabile quantitativa sulla base di una variabile qualitativa (fattore su più livelli).



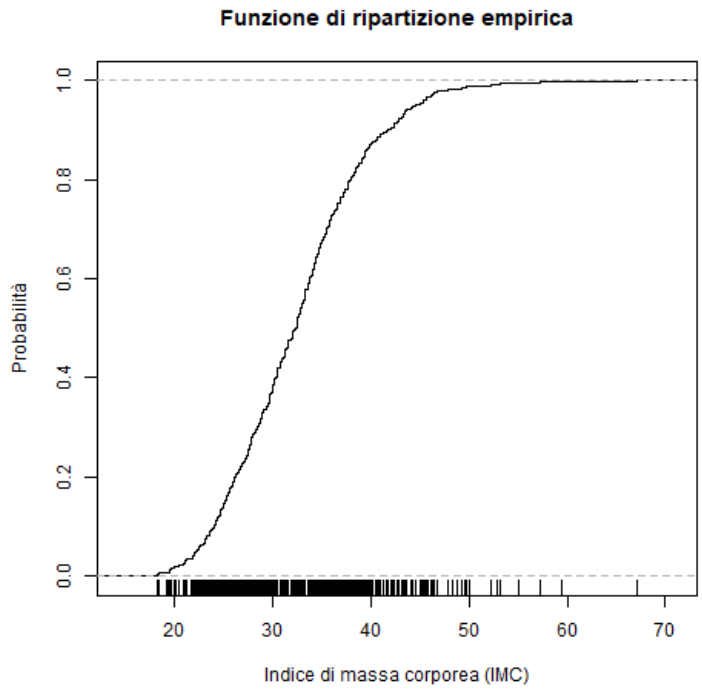


Nel primo grafico notiamo che in generale le donne con il diabete tendono ad avere un IMC più elevato rispetto a quella senza; le mediane sono rispettivamente 34.3 e 30.1 (in entrambi i casi valori molto elevati). Per il secondo grafico, condizionato alle classi d'età vediamo che, più o meno, tutte le classi si trovano nelle stesse condizioni, ossia mediana simile e range del 50% delle osservazioni simile.

Forse le categorie 21-25 e Over 50 tendono ad avere valori più bassi dell'IMC rispetto alle altre classi (ovviamente non sappiamo ancora se queste differenze sono significative oppure no).

Funzione di ripartizione empirica

La funzione di ripartizione empirica è una statistica “distribution-free” che fornisce la percentuale di osservazioni campionarie minori o uguali ad un determinato valore. Il grafico fornito da R riguardo la variabile quantitative IMC è il seguente:



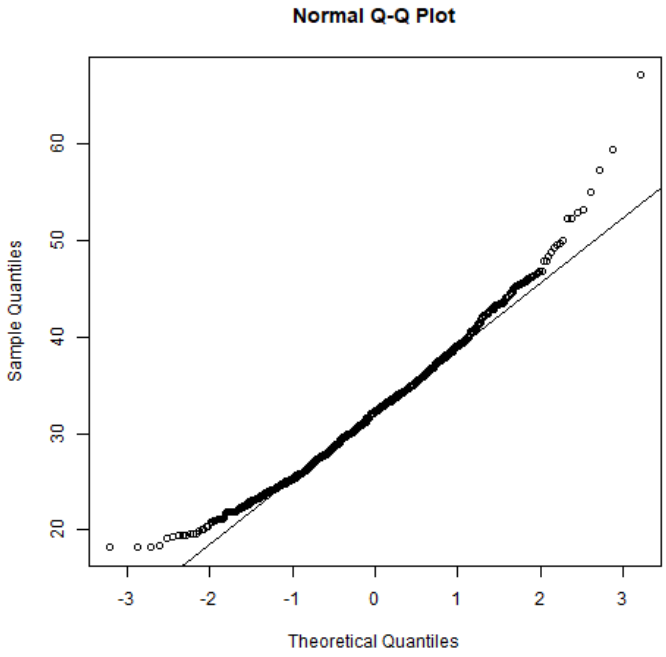
Notiamo che la funzione empirica della variabile IMC potrebbe non discostarsi molto dalla funzione empirica di una variabile normale, forse l’unica cosa che non ci da questa sicurezza è la coda destra che scende molto lentamente (a causa di valori anomali elevati). Inoltre, è possibile vedere la mediana anche da questo grafico, proiettando sull’asse delle x il valore della $y = 0.5$ (risulta, come avevamo già visto 32.30).

Massima verosimiglianza

La variabile IMC potrebbe essere attribuita ad una variabile casuale normale, quindi, supponendo l'estrazione del campione proprio da tale variabile possiamo calcolare le stime, con il metodo della massima verosimiglianza, dei parametri μ e σ^2 che risultano rispettivamente la media e la varianza stimate sul campione, ossia $\bar{x}$ e s^2.

```
> mean(BMI)
[1] 32.45746
> var(BMI)
[1] 47.95546
```

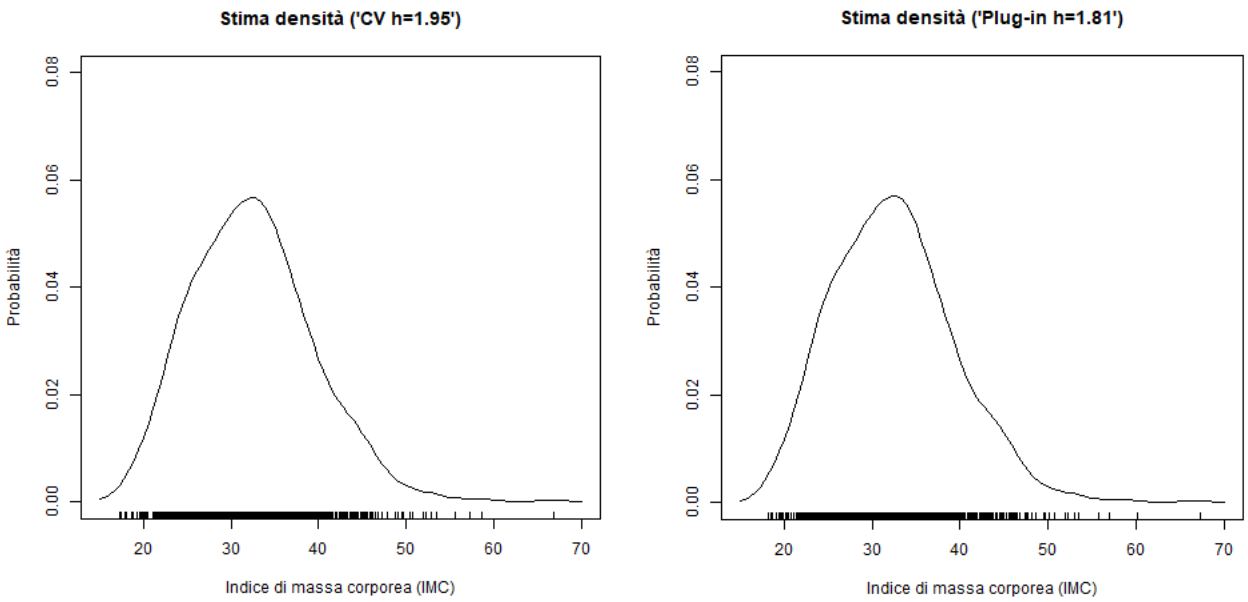
Inoltre, tramite un qqplot, che mette a confronti i quantili empirici della distribuzione con quelli di una normale standardizzata, possiamo andare a vedere graficamente se questa normalità è rispettata oppure no.



Dal qqplot non è molto evidente la normalità, anzi, fa tendere l'ago della bilancia verso una distribuzione che non è una normale.

Stimatori di nucleo univariati

Gli ‘stimatori di nucleo’ vanno a stimare la funzione di densità di una variabile quantitativa, in questo caso la variabile IMC; dipendono da un parametro `h`, chiamato ‘parametro di smorzamento’ che rende la funzione più (o meno) liscia. Questo parametro può essere scelto in modo arbitrario oppure automaticamente in base a 2 metodi, ossia la stima *Cross-Validation* e la stima *Plug-in*.

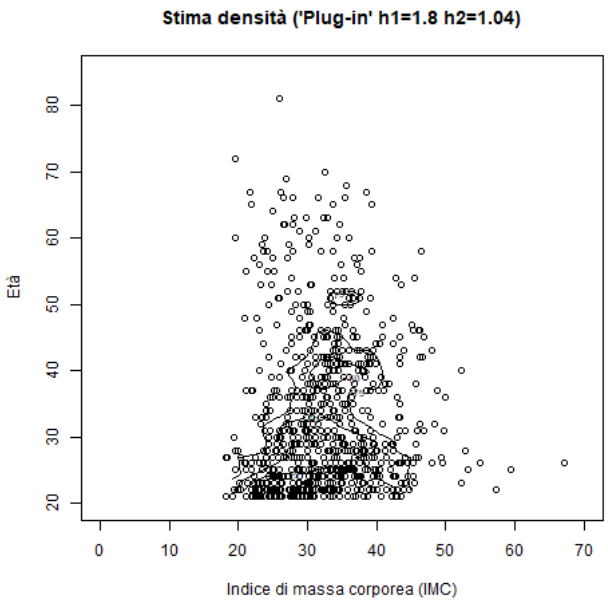


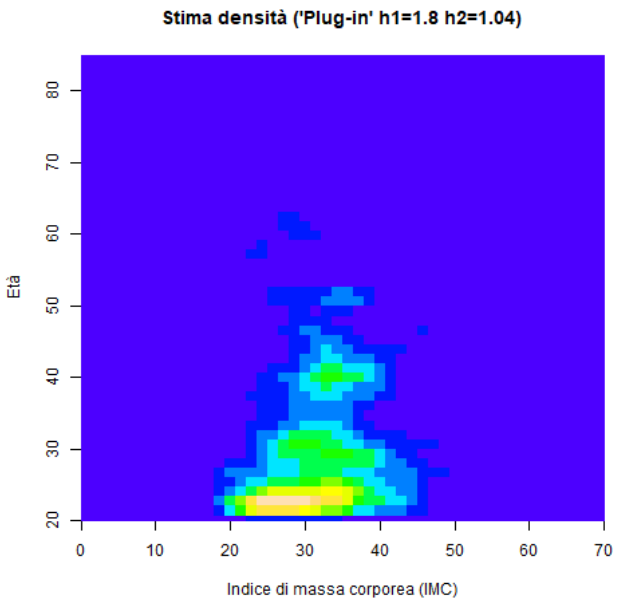
Vediamo che le stime ottenute con i 2 metodi sono pressoché identiche poiché il parametro `h` stimato è molto simile (1.95 con *CV*, 1.81 con *Plugin*); la funzione di densità stimata ha la classica funziona campanulare però si vede molto bene la leggera asimmetria positiva dovuta ai valori anomali.

Stimatori di nucleo bivariati

Lo stesso stimatore può essere applicato anche al mondo bivariato, in questo caso la coppia di variabili IMC e Age, andando a creare modelli 3D, oppure modelli che rappresentano la terza dimensione con le curve di livello (metodo adottato qui).

Per quanto riguarda il parametro di smorzamento `h`, è stato scelto di stimarlo attraverso il metodo *Plugin*.





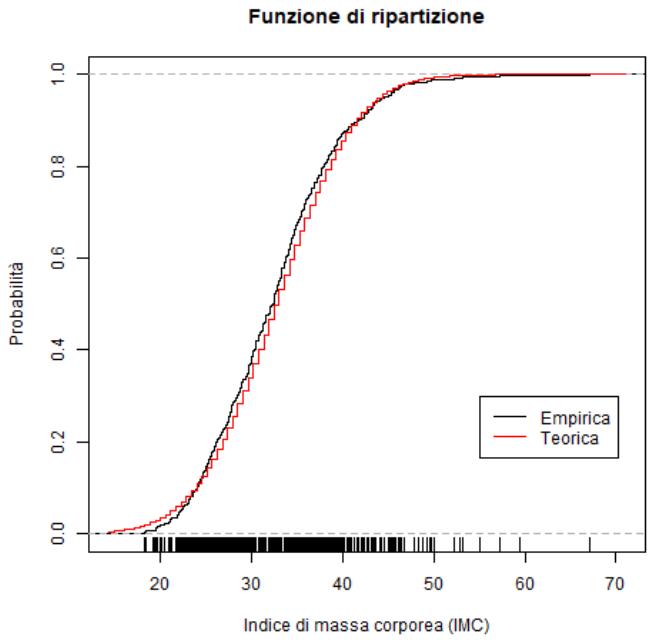
Anche da qui non risulta nessuna dipendenza lineare, però sembrerebbe che si siano formati più ‘cluster’, il primo (il principale) a ridosso dei 21-25 anni con IMC che oscilla maggiormente tra 20 e 40, il secondo verso i 30 anni che invece incorpora IMC tra 25 e 40e infine, un cluster creatosi a ridosso dei 40 anni che oscilla tra IMC = 30 e IMC = 40.

Inoltre, in modo meno evidente, si crea anche un piccolo cluster verso i 50 anni.

Analisi inferenziale

Test di Kolmogorov

Dopo un'analisi esplorativa univariata e bivariata si è passati ad un'analisi inferenziale, facendo delle ipotesi. La prima ipotesi che è stata fatta si riferiva non ad un determinato parametro ma all'intera funzione di ripartizione. L'ipotesi di base H_0 prevede che la funzione di ripartizione empirica sia uguale ad una funzione di ripartizione teorica (nel nostro caso abbiamo scelto una distribuzione normale), quindi $H_0: F(x)=F_0(x)$, contro l'ipotesi alternativa $H_1: F(x)\neq F_0(x)$. Quest'ipotesi è chiamata 'Test di Kolmogorov':



La funzione di ripartizione teorica (ossia la normale) si 'fitta' molto bene a quella empirica, solamente che quest'ultima ha la coda destra più lunga (più lenta a scendere) e quella sinistra più corta (scende troppo velocemente) rispetto alla normale.

Implementando il test:

```
> ks.test(BMI, "pnorm", 32.46, 6.92)

        One-sample Kolmogorov-Smirnov test

data:  BMI
D = 0.035105, p-value = 0.3084
alternative hypothesis: two-sided
```

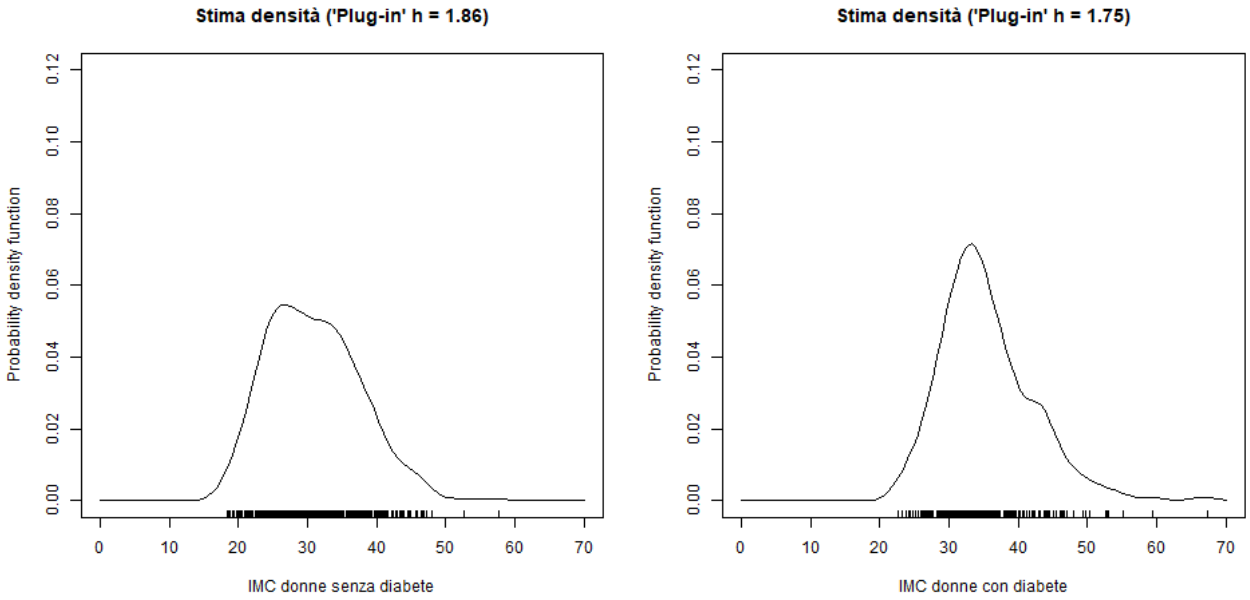
Abbiamo che il *p-value* > 0.05, quindi accettiamo l'ipotesi di base H_0 che suppone l'uguaglianza delle due funzioni di ripartizione.

Test T per due campioni

Il Test T-Student suddivide una variabile quantitativa, IMC, di numerosità n (in base ad un fattore a 2 livelli, in questo caso Diabete) in 2 variabili di numerosità n_1 e n_2, con $n_1 + n_2 = n$.

Il test fa delle assunzioni molto forti, come l'ipotesi di normalità dei 2 sotto campioni e l'ipotesi di 'omoschedasticità' (uguaglianza delle varianze), per poi andare a verificare l'omogeneità delle medie; si ha quindi $H_0{:}\mu_1 = \mu_2$, contro $H_1{:}\mu_1 \neq \mu_2$ (una ipotesi bilaterale) oppure $H_0{:}\mu_1 = \mu_2$, contro $H_1{:}\mu_1 > \mu_2$ ($H_0{:}\mu_1 = \mu_2$, contro $H_1{:}\mu_1 < \mu_2$, ipotesi direzionale). In questo caso è stata fatta un'ipotesi direzionale del tipo $H_0{:}\mu_1 = \mu_2$, contro $H_1{:}\mu_1 < \mu_2$.

Prima di andare ad implementare il test per l'uguaglianza delle medie conviene prima analizzare le 2 ipotesi forti che sono state fatte, la prima attraverso la stima di nucleo dei 2 sotto-campioni e la seconda attraverso un test che va a verificare l'uguaglianza delle varianze.



Dai grafici la normalità non è molto evidente, bisogna anche considerare però che la numerosità campionaria è più bassa adesso; per quanto riguarda l'uguaglianza delle varianze bisogna effettuare un apposito test basato sulle ipotesi H_0: $\sigma_1^2 = \sigma_2^2$ e H_1: $\sigma_1^2 \neq \sigma_2^2$. Il test fa il rapporto delle varianze ed è chiamato 'Test di Fisher'.

```
> var.test(BMI_NO, BMI_SI, paired = F, alternative = "two.sided")

        F test to compare two variances

data:  BMI_NO and BMI_SI
F = 0.98367, num df = 490, denom df = 265, p-value = 0.8702
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7932778 1.2115225
sample estimates:
ratio of variances
         0.9836665
```

Il test di fornisce un *p-value* molto alto, quindi l'ipotesi H_0: $\sigma_1^2 = \sigma_2^2$ viene accettata e possiamo implementare il Test T-Student.

```
> t.test(BMI_NO, BMI_SI, var.equal = T, alternative = "less")

        Two Sample t-test

data:  BMI_NO and BMI_SI
t = -8.4718, df = 766, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf -3.89781
sample estimates:
mean of x mean of y
 30.30420  35.14254
```

Il *p-value=0* ci fa accettare l'ipotesi H_1:$\mu_1 < \mu_2$ dandoci anche delle stime relative a μ_1 e μ_2, rispettivamente 30.86 e 35.41.

Test di Mann-Whitney

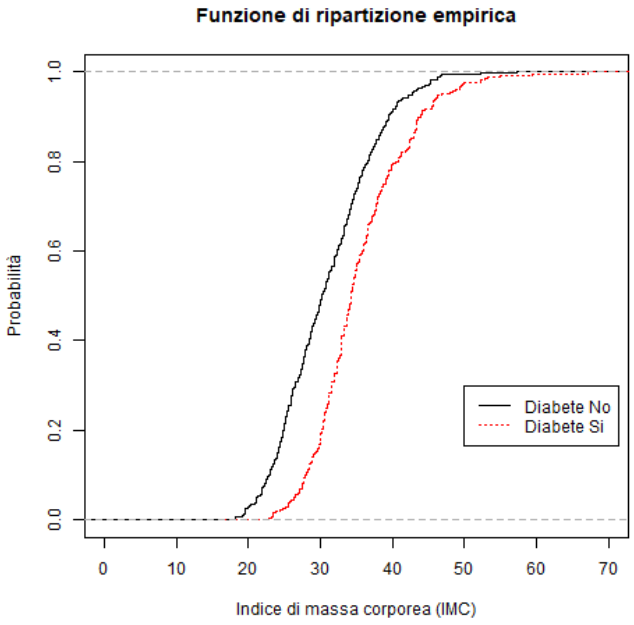
Il Test di Mann-Whitney fa sostanzialmente la stessa cosa del Test T, solamente che qui si ha un approccio 'distribution free' e quindi il test è centrato sulla mediana (stimatore più robusto che non è influenzato dai valori anomali) e non sulla media, quindi si avrà $H_0{:}\lambda_1 = \lambda_2$ contro $H_1{:}\lambda_1 \neq \lambda_2$(ipotesi bilaterale) oppure $H_0{:}\lambda_1 = \lambda_2$, contro $H_1{:}\lambda_1 > \lambda_2$ ($H_0{:}\lambda_1 = \lambda_2$, contro $H_1{:}\lambda_1 < \lambda_2$ipotesi direzionale). Le variabili prese in esame sono sempre IMC e Diabete:

```
> wilcox.test(BMI ~ Diabetes, alternative = "two.sided")

        Wilcoxon rank sum test with continuity correction

data:  BMI by Diabetes
W = 40875, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Il *p-value* praticamente =0 ci fa rifiutare $H_0{:}\lambda_1 = \lambda_2$, quindi le 2 mediane sono statisticamente differenti, con $\lambda_1 < \lambda_2$ (dove λ_1 è la mediana delle donne senza diabete e λ_2 la mediana delle donne con diabete).

Test di Kolmogorov-Smirnov

Questo 'Test di Kolmogorov-Smirnov' è incentrato sulle funzioni di ripartizione, in particolar modo si suddivide una variabile quantitativa (nel nostro caso IMC) in 2 sotto campioni in base ad un fattore a 2 livelli (nel nostro caso Diabete) e si fa l'ipotesi di uguaglianza delle funzioni di ripartizione dei 2 sotto campioni, ossia $H_0{:}F_{c1}(y) = F_{c2}(y)$ contro $H_1{:}F_{c1}(y) \neq F_{c2}(y)$.




```
> ks.test(BMI_NO, BMI_SI)

        Two-sample Kolmogorov-Smirnov test

data:  BMI_NO and BMI_SI
D = 0.30944, p-value = 8.993e-15
alternative hypothesis: two-sided
```

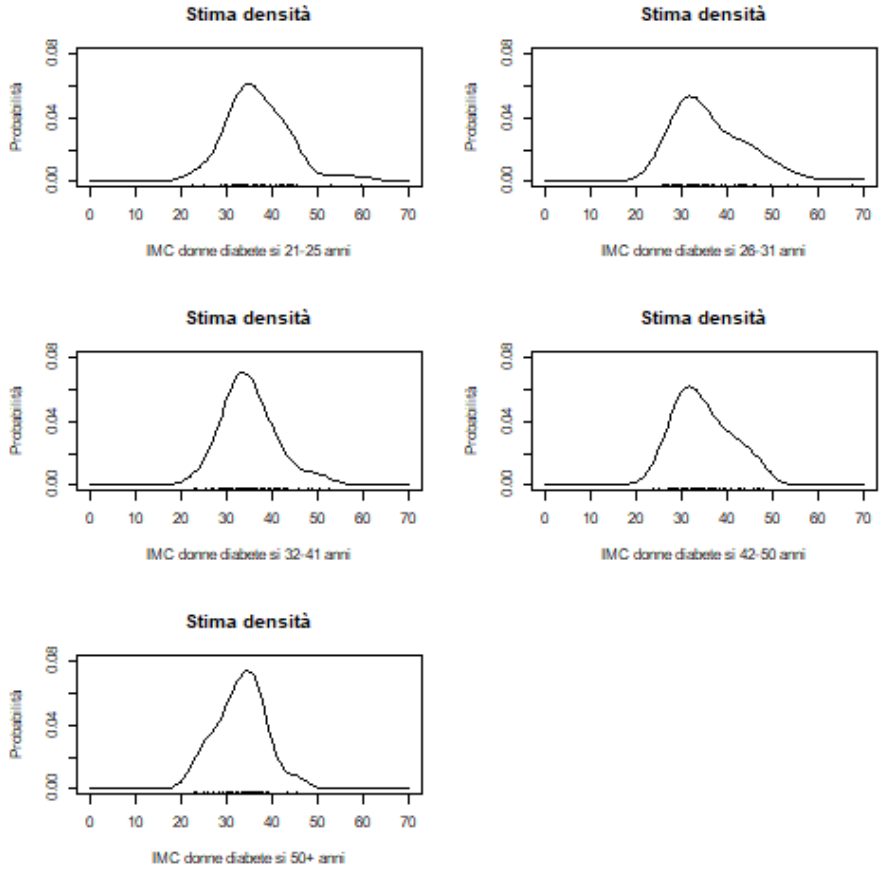
Già dal grafico notiamo che i valori di IMC quando le donne hanno il diabete sono nettamente più elevati, infatti le funzioni di ripartizioni sono nettamente distaccate ed infatti il test conferma che le due funzioni sono statisticamente differenti con un *p-value* molto basso che fa accettare l'ipotesi H_1:$F_{c1}(y) \neq F_{c2}(y)$.

Analisi della varianza

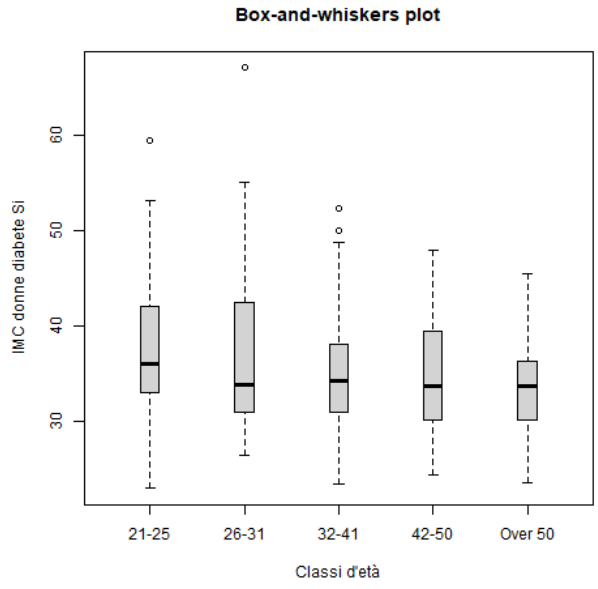
L'analisi della varianza può essere vista come un'estensione del Test T, poiché fa esattamente le stesse ipotesi e assunzioni, l'unica differenza è che non si ha un fattore a 2 livelli ma un fattore a *'r'* livelli (nel nostro caso Age_Class, 5 livelli).

Il sistema di ipotesi è quindi H_0:$\mu_1 = \mu_2 = \cdots = \mu_r$ contro H_1: $\mu_j \neq \mu_l$ (almeno una media differisce dalle altre). La variabile quantitativa in questo caso è l'IMC relativo solamente alle donne con diabete.

Per quanto riguarda la normalità facciamo le stime di nucleo dei 5 livelli:



Dalle stime di nucleo tutte e 5 le distribuzioni sembrerebbero non distaccarsi troppo da una distribuzione normale. Si potrebbe andare a vedere anche un boxplot:



Dal boxplot vediamo che la classe delle donne con diabete con IMC più alto è quella 21-25 anni; ciò che ci interessa però è vedere se queste medie differiscono statisticamente oppure queste differenze sono dovute semplicemente al campione. Implementiamo il test.

```
> summary(aov(Data_Diab_Si$BMI ~ Data_Diab_Si$Age_Class))
                          Df Sum Sq Mean Sq F value Pr(>F)
Data_Diab_Si$Age_Class     4    435  108.82   2.545   0.04 *
Residuals                261  11161   42.76
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Notiamo che il *p-value* è 0.04, quindi nel caso di significatività al 5% si rifiuta H_0 e si accetta $H_1: \mu_j \neq \mu_l$, quindi c'è almeno una media che differisce statisticamente dalle altre. Per vedere qual è questa, si considera la 'procedura di Tukey' che crea degli intervalli di confidenza simultanei per ogni coppia:

```
> TukeyHSD(aov(Data_Diab_Si$BMI ~ Data_Diab_Si$Age_Class), "Data_Diab_Si$Age_Class")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Data_Diab_Si$BMI ~ Data_Diab_Si$Age_Class)

$`Data_Diab_Si$Age_Class`
                    diff       lwr        upr     p adj
26-31-21-25 -0.3243275 -3.906360 3.25770503 0.9991516
32-41-21-25 -1.9997661 -5.378490 1.37895821 0.4822266
42-50-21-25 -2.2194771 -5.893293 1.45433881 0.4609302
51+-21-25   -3.8993393 -7.885666 0.08698774 0.0586084
32-41-26-31 -1.6754386 -4.822859 1.47198154 0.5879424
42-50-26-31 -1.8951496 -5.357432 1.56713273 0.5610680
51+-26-31   -3.5750119 -7.367278 0.21725386 0.0752280
42-50-32-41 -0.2197110 -3.471206 3.03178402 0.9997334
51+-32-41   -1.8995733 -5.500420 1.70127388 0.5964775
51+-42-50   -1.6798622 -5.558940 2.19921575 0.7573882
```

Vediamo che tutte le coppie hanno praticamente *p-value >0.05*, quindi non ci sono delle grosse differenze tra le varie coppie, infatti tutti gli intervalli contengono lo 0, al massimo le coppie 51+/21-25 e 51+/26-31 hanno i valori più bassi del *p-value*, ma comunque più elevati di 0.05. In questo caso ancora non si può concludere nulla, a meno che non si consideri un livello di significatività 0.01 (consigliabile in ambito medico) che ci porta ad accettare $H_0: \mu_1 = \mu_2 = \cdots = \mu_r$, ossia nessuna differenza tra i gruppi.

Test di Kruskal-Wallis

Probabilmente per vedere se c'è differenza significativa tra i vari gruppi è conveniente passare ad un approccio 'distribution-free' piuttosto che un 'approccio classico' poiché non si fanno assunzioni. Il Test di Kruskal-Wallis si comporta come quello di Mann-Whitney, solamente che si ha un fattore a più livelli (in questo caso Age_Class con 5 livelli); il sistema di ipotesi è composto così, $H_0{:}\lambda_1 = \lambda_2 = \cdots = \lambda_r$ contro $H_1{:}\lambda_j \neq \lambda_l$. Le variabili considerate sono sempre IMC (solo relative alle donne con diabete) e l'età in classi. Il test implementato in R dà il seguente risultato:

```
> kruskal.test(Data_Diab_Si$BMI ~ Data_Diab_Si$Age_Class)

        Kruskal-Wallis rank sum test

data:  Data_Diab_Si$BMI by Data_Diab_Si$Age_Class
Kruskal-Wallis chi-squared = 7.1632, df = 4, p-value = 0.1275
```

Qui non ci sono molto dubbi, in quanto il *p-value* è abbastanza alto e si accetta $H_0{:}\lambda_1 = \lambda_2 = \cdots = \lambda_r$, quindi non la differenza tra le mediane dei gruppi non sono statisticamente significative.

Regressione logistica

La 'regressione logistica' è un caso particolare del modello lineare generalizzato in cui la variabile di risposta *'Y'* è una variabile binaria e quindi si assume che *'Y'* è una variabile casuale di Bernoulli. In questo caso è stata fatta una regressione logistica usando come variabile di risposta la variabile dicotomica Diabete (che assume valori "Si" e "No") e regressore la variabile quantitativa IMC.

Costruendo tale modello in R avremo un output:

```
> model<-glm(Diabete ~ IMC,data=Data, family=binomial)
> summary(model)

Call:
glm(formula = Diabete ~ IMC, family = binomial, data = Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0094  -0.9184  -0.6598   1.2254   1.9107

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.99682    0.42885   -9.32  < 2e-16 ***
IMC          0.10250    0.01261    8.13 4.31e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 981.53  on 756  degrees of freedom
Residual deviance: 904.89  on 755  degrees of freedom
AIC: 908.89

Number of Fisher Scoring iterations: 4
```

Dall'output vediamo che, sia l'intercetta e il coefficiente di IMC sono significativi ma nel complesso il modello non sembra dei migliori, infatti si ha un'alta devianza residua che indicano una grande dispersione. Anche l'AIC è molto elevato. Volendo vedere più nello specifico l'impatto del modello costruiamo la 'probabilità di avere il diabete' per ogni unità statistica.

```
> model.prob<-predict(model,type="response")
> probabib<-round(model.prob,digits=4)
> dffit<-cbind(Data,"Pr. Diabete"=probabib)
> head(dffit)
  Diabete  IMC Età Classi_età Pr. Diabete
1      Si 33.6  50      42-50      0.3652
2      No 26.6  31      26-31      0.2192
3      Si 23.3  32      32-41      0.1668
4      No 28.1  21      21-25      0.2466
5      Si 43.1  33      32-41      0.6037
6      No 25.6  30      26-31      0.2022
```

Per le prime 6 unità statistiche del campione abbiamo i rispettivi valori delle variabili con la relativa probabilità di avere il diabete; per queste unità vediamo che il modello sbaglia 1 volta su 6, ossia nel caso dell'osservazione 3 che il modello prevede abbia il diabete ma in realtà non è così (non sono da considerare le altre variabili in quanto quelle utilizzate nel modello sono solamente Diabete e IMC).

Sulla base di quanto detto andiamo a vedere quante volte il nostro modello ha predetto bene (oppure no):

```
> model.pred=rep("No",nrow(Data))
> model.pred[model.prob>0.5]="Si"
> table(model.pred,Data$Diabete)

model.pred  No  Si
        No 432 199
        Si  59  67
```

Dalla tabella vediamo che il modello ha predetto bene se una persona avesse (o meno) il diabete in 499 casi su 757 (nel 65% dei casi). Per le persone senza diabete ha predetto correttamente 432 casi su 491 (88% dei casi), mentre per le persone con diabete solamente 67 casi su 266 (solo 25%).

In seguito, si è voluto vedere se il modello migliorasse con l'aggiunta del regressore Age_Class.

```
> model2<-glm(Diabete ~ IMC + Classi_età,data=Data, family=binomial)
> summary(model2)

Call:
glm(formula = Diabete ~ IMC + Classi_età, family = binomial,
    data = Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8678  -0.8804  -0.5088   1.0613   2.3033

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.94929    0.48470 -10.211  < 2e-16 ***
IMC             0.10349    0.01334   7.757 8.68e-15 ***
Classi_età26-31  0.69399    0.24375   2.847  0.00441 **
Classi_età32-41  1.49487    0.24024   6.223 4.89e-10 ***
Classi_età42-50  1.66883    0.27939   5.973 2.33e-09 ***
Classi_età51+    1.60916    0.29446   5.465 4.63e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 981.53  on 756  degrees of freedom
Residual deviance: 837.89  on 751  degrees of freedom
AIC: 849.89

Number of Fisher Scoring iterations: 4
```

In effetti sembra che un miglioramento ci sia dato che sia la devianza dei residui che l'AIC sono più bassi rispetto al modello precedente.

```
> model2.probs<-predict(model2,type="response")
> probabib2<-round(model2.probs,digits=4)
> dffit2<-cbind(Data,"Pr. Diabete"=probabib2)
> head(dffit2)
  Diabete  IMC Età Classi_età Pr. Diabete
1      Si 33.6  50      42-50      0.5490
2      No 26.6  31      26-31      0.1820
3      Si 23.3  32      32-41      0.2605
4      No 28.1  21      21-25      0.1149
5      Si 43.1  33      32-41      0.7322
6      No 25.6  30      26-31      0.1671
```

Costruendo le probabilità per ogni osservazione notiamo che, nei primi 6 casi, la situazione non cambia, per questo conviene andare a vedere direttamente la situazione generale.

```
> model2.pred=rep("No",nrow(Data))
> model2.pred[model2.probs>0.5]="Si"
> table(model2.pred,Data$Diabete)

model2.pred  No  Si
         No 413 147
         Si  78 119
```

In questo il modello ha previsto correttamente 532 casi su 757 (70% dei casi). Riguardo alle donne senza diabete ha previsto bene 413 casi su 491 (84% dei casi), mentre riguardo le donne con diabete 119 casi su 266 (45% dei casi).

Rispetto al modello di prima l'efficacia di previsione è aumentata nel complesso da 65% a 70%; quella riferita alle donne senza diabete è diminuita da 88% a 84% ma soprattutto quella relativa alle donne con diabete è aumentata da 25% a 45%. Il modello prevede correttamente se una donna ha il diabete quasi 1 volta su 2 (ovviamente è ancora una previsione bassa ma è un miglioramento rispetto all'altro modello). Con qualche altra variabile (regressore) a disposizione si potrebbero probabilmente modelli migliori che prevedano in modo più efficace.