Progetto FSM - dataset birthwt

Dario Cioni

11/02/2022

Contents

1 Introduzione

L'obiettivo del presente elaborato è studiare i fattori che contribuiscono al basso peso alla nascita nei neonati, analizzando i dati presenti nel dataset birthwt.
Il dataset contiene dati raccolti su 189 bambini nati al Baystate Medical Center, Springfield, Mass nel 1986. Le variabili coinvolte sono 10:

- bwt: peso alla nascita espresso in grammi
- age: età della madre
- lwt: peso della madre (espresso in libbre) alla fine dell'ultimo periodo mestruale
- race: etnia della madre (1=bianca, 2=nera, 3=altro)
- smoke: indica se la madre è fumatrice, (1=fumatrice,0 altrimenti)
- ptl: numero di precedenti parti prematuri
- ht: indica se esiste una storia di ipertensione (1=presente,0 assente)
- ui: indica la presenza di irritabilità uterina (1=presente,0 assente)
- ftv: numero di visite dal ginecologo nel primo trimestre
- low: variabile dicotomizzata da bwt, indica se il bambino è al di sotto di 2.5 kg

2 Analisi esplorativa

Le variabili presenti nel dataset possono essere suddivise nel seguente modo

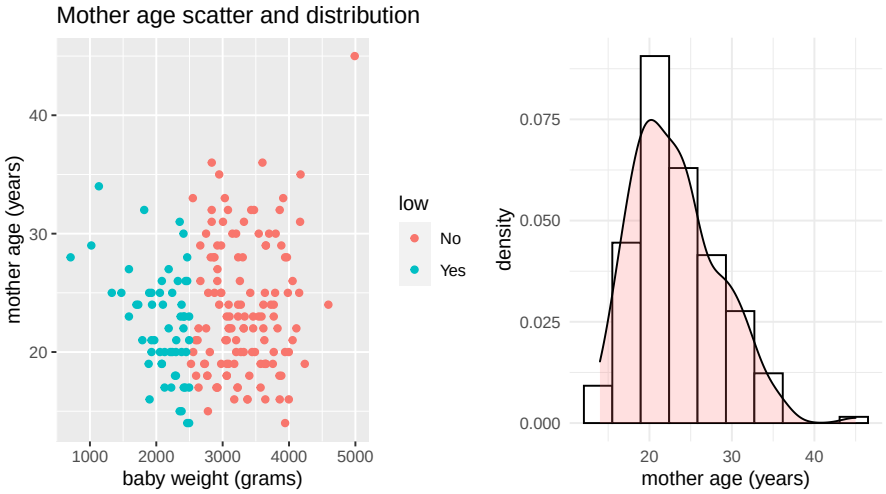
- due variabili a valori continui (bwt,lwt)
- tre variabili a valori discreti (age,ptl,ftv)
- cinque variabili categoriche (race,smoke,ht,ui,low)

```
##    low age lwt  race smoke ptl ht  ui ftv  bwt
## 85  No  19 182 Black    No   0 No Yes   0 2523
## 86  No  33 155 Other    No   0 No  No   3 2551
## 87  No  20 105 White   Yes   0 No  No   1 2557
## 88  No  21 108 White   Yes   0 No Yes   2 2594
## 89  No  18 107 White   Yes   0 No Yes   0 2600
## 91  No  21 124 Other    No   0 No  No   0 2622
##   low          age            lwt           race    smoke     ptl
##  No :130   Min.   :14.00   Min.   : 80.0   White:96   No :115   0:159
##  Yes: 59   1st Qu.:19.00   1st Qu.:110.0   Black:26   Yes: 74   1: 24
```

```
##            Median :23.00   Median :121.0   Other:67            2:  5
##            Mean   :23.24   Mean   :129.8                       3:  1
##            3rd Qu.:26.00   3rd Qu.:140.0
##            Max.   :45.00   Max.   :250.0
##    ht        ui      ftv          bwt
##  No :177   No :161   0:100   Min.   : 709
##  Yes: 12   Yes: 28   1: 47   1st Qu.:2414
##                      2: 30   Median :2977
##                      3:  7   Mean   :2945
##                      4:  4   3rd Qu.:3487
##                      6:  1   Max.   :4990
```

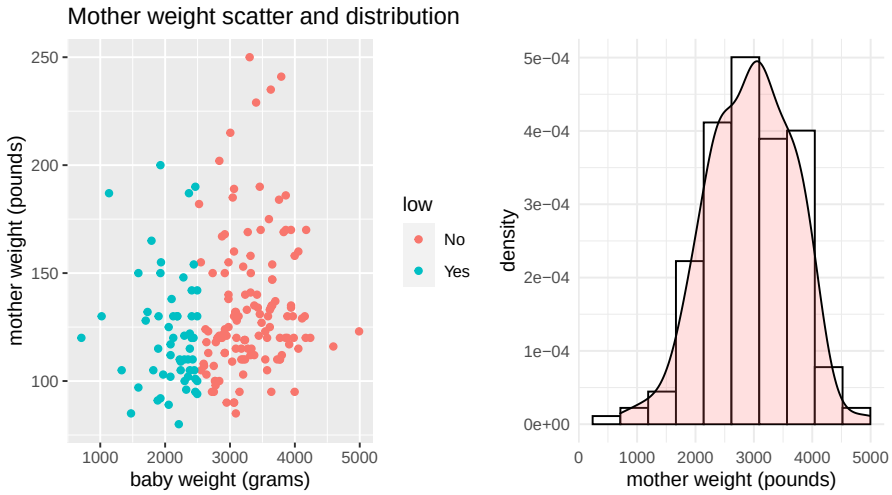
2.1 Variabili continue

Attraverso dei grafici è possibile studiare la composizione del dataset e vedere se è presente una relazione tra le variabili continue la variabile obiettivo bwt. Studiamo la distribuzione dei dati mediante degli scatter plot

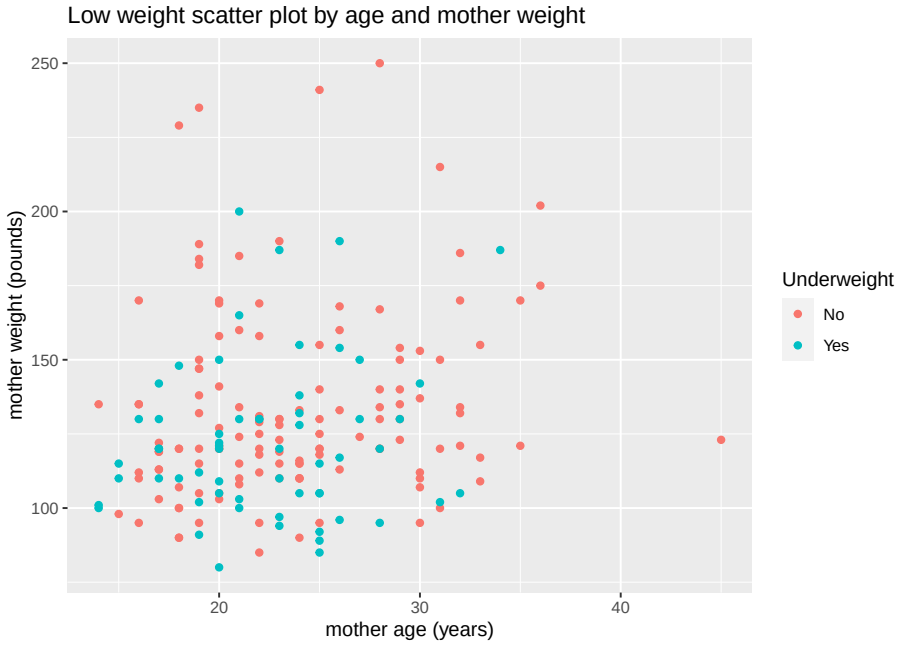


L'età delle madri va da 14 a 45 anni, con mediana di 23 anni. Possiamo notare che la distribuzione è spostata verso sinistra. Un modo di normalizzarla è l'utilizzo della radice quadrata dell'età

```
birthwt$age <- sqrt(birthwt$age)
```



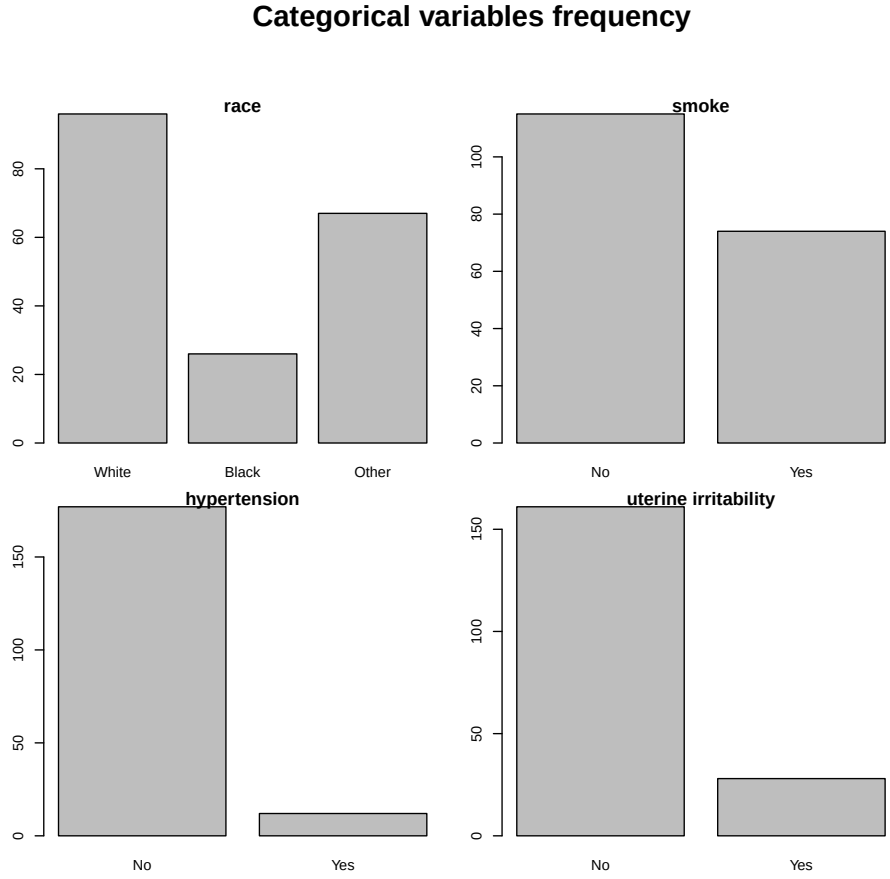
Il peso della madre all'ultimo ciclo mestruale va da 80 libbre a 250 libbre (da 36 a 113 kg), con mediana 121 (55 kg).
In questo caso la distribuzione non risulta squilibrata.



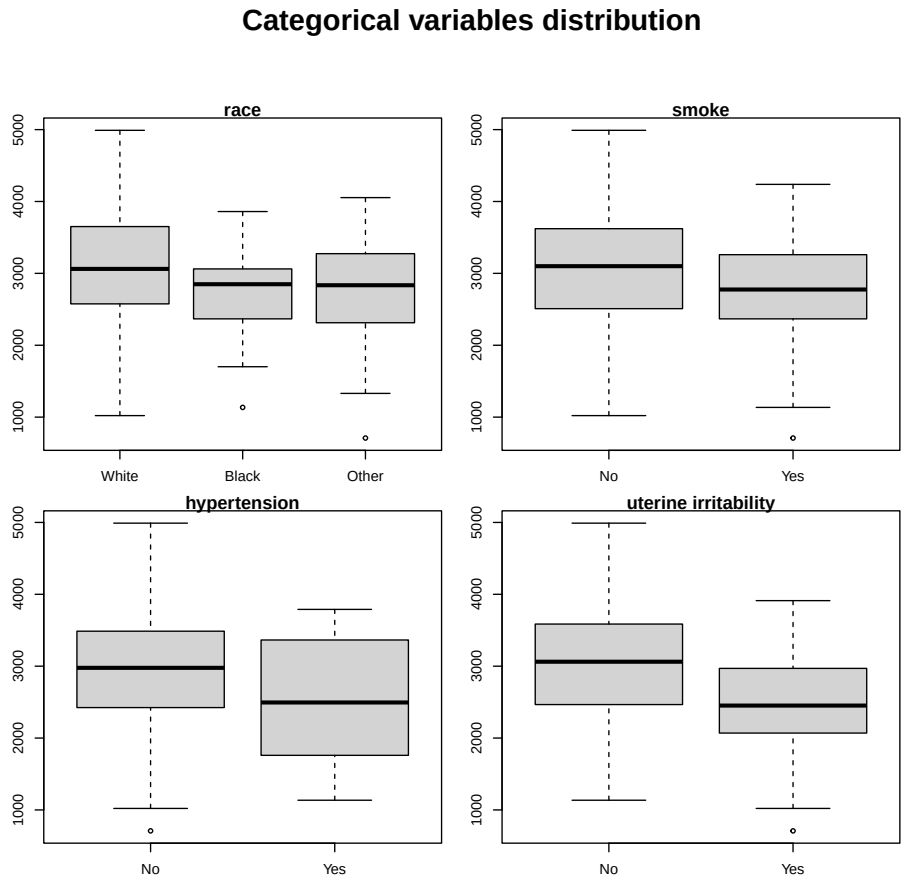
Gli scatter plot non evidenziano una chiara relazione, lineare o polinomiale, tra le variabili age e lwt e la variabile di uscita bwt, nè una chiara interazione tra age e lwt.

2.2 Variabili categoriche

Le variabili categoriche smoke, ht e ui sono binarie. Per smoke è presente un buon numero di campioni, mentre le madri con ipertensione e irritabilità uterina sono presenti in numero più ridotto nel dataset. La variabile race possiede tre valori, White, Black e Other. Questa suddivisione è abbastanza grossolana, ed il campioni di madri nere è inferiore alle altre due categorie.



Attraverso dei boxplot è possibile studiare se esiste un legame tra queste variabili e la variabile risposta bwt.

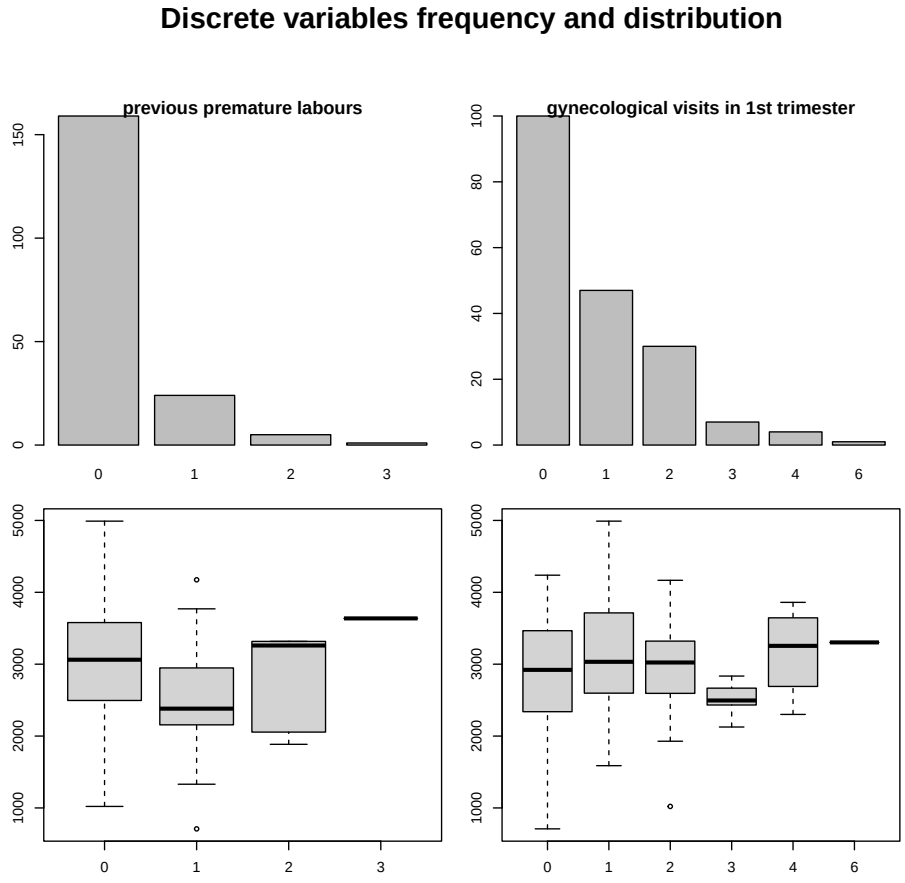


- Le madri madri nere o di altra etnia hanno mediana del peso dei bambini inferiore rispetto alle madri bianche.
- Nelle madri fumatrici la mediana di bwt è inferiore
- Nel caso in cui sia presente ipertensione o irritabilità uterina si ha una forte diminuzione della mediana di bwt.

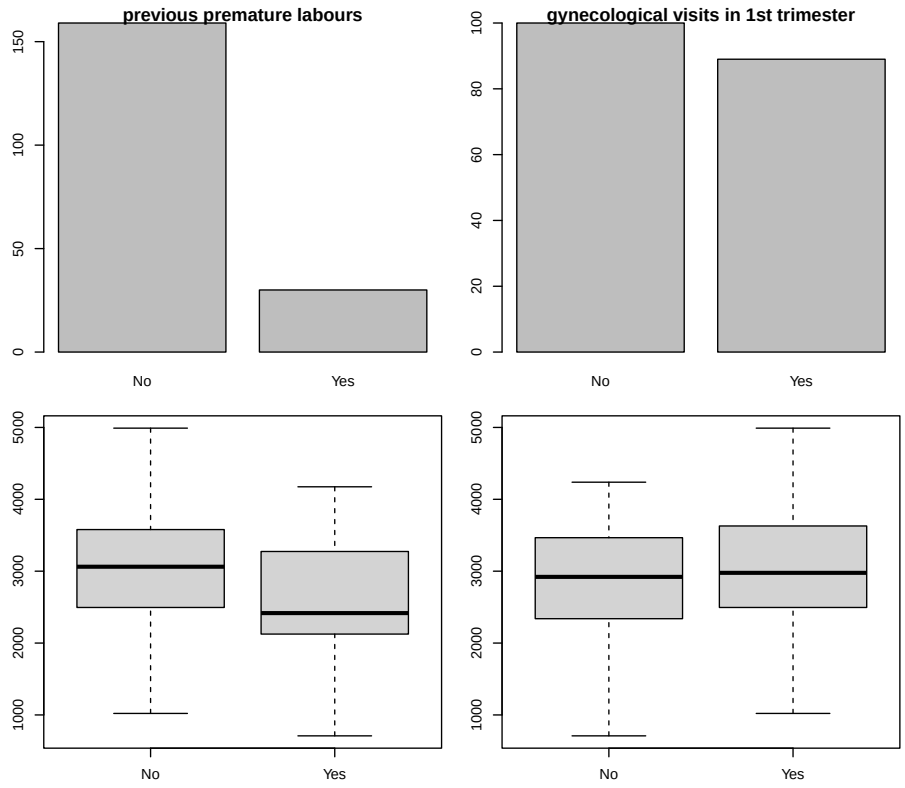
Queste 4 variabili risultano possibilmente connesse al peso e potranno essere analizzate più a fondo in seguito utilizzando modelli statistici.

2.3 Variabili discrete

Studiando la distribuzione delle variabili discrete ptl ed ftv



Le variabili ptl e ftv mostrano un elevato numero di campioni con il valore pari a zero, mentre pochi campioni all'aumentare del valore della variabile. Per sopperire a questo squilibrio, potranno essere dicotomizzate in variabili binarie, dove 0 equivale all'assenza della caratteristica, ed 1 equivale alla presenza della caratteristica in valore maggiore o uguale a 1.



Le madri che hanno avuto precedenti parti prematuri, sembrano avere una mediana del peso inferiore: pur avendo pochi dati a disposizione possiamo pensare che esista una relazione tra il numero di parti prematuri della madre e il peso del nascituro. Il numero di visite dal ginecologo nel primo trimestre non sembra essere così rilevante, in quanto evidenzia una differenza minima nelle madri che non hanno effettuato visite rispetto alle madri che hanno fatto almeno una visita.

2.4 Analisi delle correlazioni

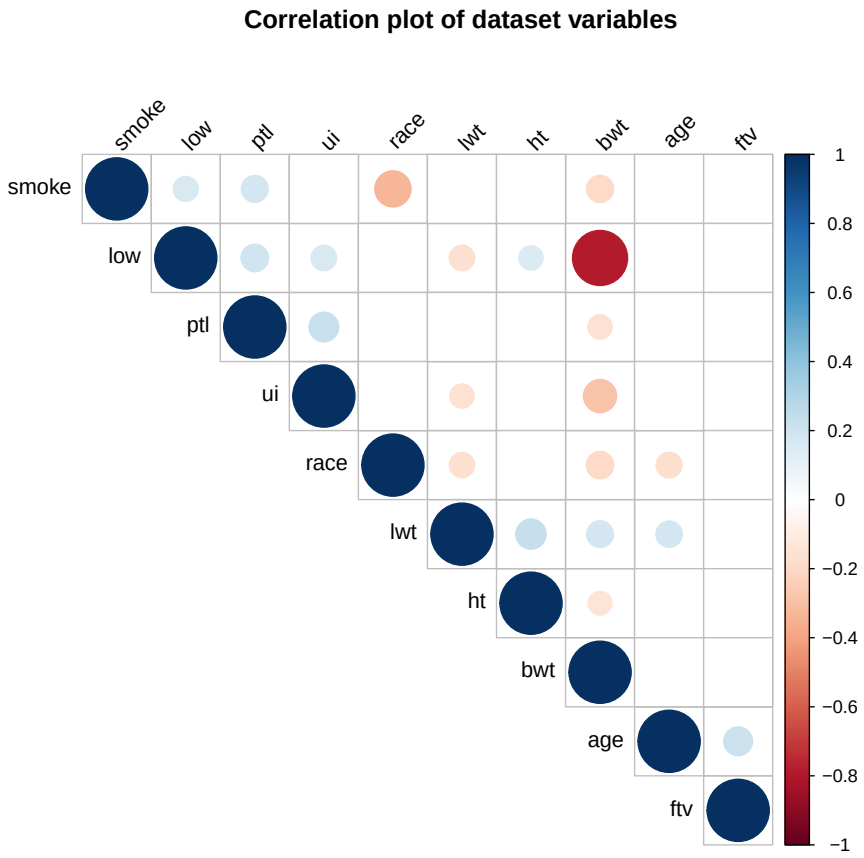
Il dataset fornito ha un buon numero delle variabili: per capire meglio le relazioni tra le variabili è possibile studiare la correlazione tra queste. In questo modo, è possibile

- Verificare se esistono variabili collineari
- Individuare possibili interazioni tra variabili

- Individuare variabili scorrelate in modo da semplificare la successiva creazione di modelli

Si calcola perciò la correlazione di Pearson ed il p-value asintotico tra ogni coppia di variabili Nel grafico sono presenti solamente le correlazioni ritenute significative, con p-value inferiore a 0.05. Sono colorate in rosso le correlazioni negative, mentre in blu correlazioni positive tra le coppie di variabili. Questa informazione potrà essere utilizzata per creare modelli con interazioni tra le variabili

```
corm <- rcorr(as.matrix(src),type="pearson")

p <- corm$P
p[is.na(p)] <- 0
```



- la variabile low è ovviamente correlata a bwt, in quanto sua dicotomiz-

zazione binaria. Nei successivi studi una delle due dovrà essere sempre esclusa.

- Rimuovendo low o bwt, non si evidenziano correlazioni così alte da far pensare a collinearità. Questa ipotesi sarà ulteriormente validata in seguito.
- race risulta negativamente correlata a smoke (le madri bianche tendono a fumare più di quelle nere o di altre origini)
- lwt risulta positivamente correlata a ht
- ptl risulta positivamente correlata a ui
- ftv risulta positivamente correlata a age
- ui risulta negativamente correlata a bwt (Le madri con irritabilità uterina tendono ad avere figli che pesano di meno). Questa relazione è particolarmente interessante poichè bwt è la variabile che vogliamo stimare. L'effettiva presenza di una relazione tra le variabili può essere ulteriormente visualizzata effettuando un test di correlazione, simile a quello utilizzato per costruire il grafico

```
cor.test(src$bwt, src$ui)

##
##  Pearson's product-moment correlation
##
## data:  src$bwt and src$ui
## t = -4.0493, df = 187, p-value = 7.518e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4100408 -0.1471608
## sample estimates:
##        cor
## -0.2839274
```

La correlazione stimata è di -0.28, e poichè il p-value è inferiore al livello di significatività 0.05, è possibile rifiutare l'ipotesi nulla di non correlazione.

2.5 Considerazioni sulle variabili

- La variabile race, oltre a rappresentare una caratterizzazione morfologica e genetica, potrebbe rappresentare un indicatore socio-economico delle madri. Tuttavia, questa suddivisione è grossolana, in quanto non è espressa la composizione della categoria "Other".

- La variabile smoke non indica da quanto tempo e con quale regolarità le madri hanno fumato, nè se hanno continuato durante tutto il corso della gravidanza.
- Il peso della madre non è direttamente collegabile allo stato di salute della madre (a differenza di indici quali il grado di obesità), ma potrebbe comunque avere una relazione con il peso del figlio.

3 Modello lineare di regressione

Supponiamo di voler prevedere il peso del neonato sulla base del valore delle variabili informative. Si utilizzerà come variabile obiettivo bwt, mentre la variabile low verrà esclusa dall'analisi.

3.1 Modello di regressione semplice

Individuiamo un modello di regressione lineare semplice, includendo una sola variabile esplicativa

$$E(\text{bwt} \mid \text{X}) = \beta_0 + \beta_1 \text{X}$$

Sulla base della correlazione individuata nel capitolo precedente, è possibile verificare se la variabile ui fornisca un buon modello

```
mq0 <- lm(bwt ~ ui,data=birthwt)
summary(mq0)

##
## Call:
## lm(formula = bwt ~ ui, data = birthwt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1895.7  -535.7    31.3   555.3  1959.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3030.70      55.25  54.852  < 2e-16 ***
## uiYes        -581.27     143.55  -4.049 7.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 701.1 on 187 degrees of freedom
## Multiple R-squared:  0.08061,Adjusted R-squared:  0.0757
## F-statistic:  16.4 on 1 and 187 DF,  p-value: 7.518e-05
```

Commenti

- La presenza di irritabilità uterina ha un effetto negativo sul peso del bambino
- La variabile ui risulta altamente significativa
- La statistica F risulta altamente significativa con p-value $7.5 \cdot 10^{-5}$
- Gli indici R^2 sono molto bassi e l'errore residuo è molto alto: questo modello, molto semplice, potrebbe non essere sufficiente a spiegare il peso dei neonati.

Ripetendo questa analisi anche per le altre variabili, si trova che

- Le variabili ht, lwt,race e smoke sono significative con p-value inferiore a 0.05, mentre age,ptl ed ftv non sono significative
- Secondo i modelli stimati, ht, race, smoke hanno un effetto negativo sul peso del neonato. Inoltre, le madri di colore o di altra provenienza hanno un effetto negativo sul peso.
- La variabile lwt ha un effetto $\hat{\beta}_1 = 4.429$ positivo
- Dicotomizzando la variabile ptl come variabile binaria, questa diventa significativa. Avere avuto precedenti parti prematuri ha un effetto negativo sul peso.
- Tutti questi modelli hanno una deviazione standard superiore a quello con la variabile ui ed un indice R^2 più basso.

3.2 Modello di regressione multipla

Utilizziamo un modello più complesso, introducendo altre variabili e verificando se porta a miglioramenti.
E' possibile testare il caso estremo del modello completo, nel quale si introducono tutte le variabili esplicative. Le variabili ptl e ftv saranno dicotomizzate in variabili binarie

```
mq2 <- lm(bwt ~ age+ lwt + race + smoke + ht + ui + ptl.f + ftv.f
          ,data=birthwt)
summary(mq2)

##
## Call:
## lm(formula = bwt ~ age + lwt + race + smoke + ht + ui + ptl.f +
##     ftv.f, data = birthwt)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1874.47  -456.34    58.35   492.57  1687.43
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3042.999    483.446   6.294 2.31e-09 ***
## age          -42.687     95.070  -0.449 0.653973
## lwt            4.186      1.715   2.441 0.015632 *
## raceBlack   -476.904    149.663  -3.187 0.001699 **
## raceOther   -333.792    116.679  -2.861 0.004729 **
## smokeYes    -323.489    108.098  -2.993 0.003157 **
## htYes       -573.498    200.687  -2.858 0.004774 **
## uiYes       -492.885    137.097  -3.595 0.000419 ***
## ptl.fYes    -201.224    136.318  -1.476 0.141665
## ftv.fYes      31.013    100.726   0.308 0.758519
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 646.6 on 179 degrees of freedom
## Multiple R-squared:  0.2515,Adjusted R-squared:  0.2138
## F-statistic: 6.682 on 9 and 179 DF,  p-value: 3.147e-08
```

Commenti

- Tutte le variabili tranne lwt (peso della madre) e ftv danno un contributo negativo al peso quando il carattere è presente
- Il peso della madre è significativo, essendo inferiore al livello di significatività 0.05
- Le variabili smoke,ht e ui risultano significative
- A differenza del modello di regressione semplice, la variabile ptl dicotomizzata non è significativa.
- Le variabili ftv ed age rimangono non significative
- L'errore standard residuo risulta più basso rispetto al modello di regresssione semplice, ma sempre elevato:
  Dato il peso medio dei neonati di 2945 grammi e un errore residuo di 646.8 grammi, l'errore percentuale è del 21%.
- Gli indici R^2 ed R^2 aggiustato sono superiori al modello di regressione semplice, ma non molto alti.
- La statistica F del modello è significativa

Multicollinearità E' possibile controllare ancora una volta l'ipotesi di non collinearità utilizzando il variance inflation factor (VIF). Uno score di 1 indica una assenza di multicollinearità, mentre uno score che si avvicina a 10 indica una forte multicollinearità

```
car::vif(mq2)
```

```
##           GVIF Df GVIF^(1/(2*Df))
## age   1.179977  1        1.086267
## lwt   1.237125  1        1.112261
## race  1.402355  2        1.088214
## smoke 1.258567  1        1.121859
## ht    1.082699  1        1.040528
## ui    1.072400  1        1.035567
## ptl.f 1.121868  1        1.059183
## ftv.f 1.142842  1        1.069038
```

In questo caso tutte le variabili hanno uno score vicino ad 1, quindi non è presente multicollinearità.

Modello ridotto Consideriamo adesso un modello ridotto, nel quale si escludono le variabili non significative

```
mq3 <- lm(bwt ~  race + smoke + ht + ui + lwt,data=birthwt)
summary(mq3)
```

```
##
## Call:
## lm(formula = bwt ~ race + smoke + ht + ui + lwt, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1842.14  -433.19    67.09   459.21  1631.03
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2837.264    243.676  11.644  < 2e-16 ***
## raceBlack   -475.058    145.603  -3.263 0.001318 **
## raceOther   -348.150    112.361  -3.099 0.002254 **
## smokeYes    -356.321    103.444  -3.445 0.000710 ***
## htYes       -585.193    199.644  -2.931 0.003810 **
## uiYes       -525.524    134.675  -3.902 0.000134 ***
## lwt            4.242      1.675   2.532 0.012198 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 645.9 on 182 degrees of freedom
## Multiple R-squared:  0.2404,Adjusted R-squared:  0.2154
## F-statistic:   9.6 on 6 and 182 DF,  p-value: 3.601e-09
```

Commenti

- Si ha un miglioramento nella significatività di tutte le variabili in base al p-value
- La statistica F è altamente significativa
- L'errore standard e gli indici R^2 rimangono simili a quelli del modello completo
- Questo modello, notevolmente più semplice, adatta ancora bene i dati

E' possibile escludere anche la variabile lwt, ottenendo un modello ancora più semplice:

```
mq4 <- lm(bwt ~  race + smoke + ht + ui,data=birthwt)
summary(mq4)

##
## Call:
## lm(formula = bwt ~ race + smoke + ht + ui, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1828.68  -452.50    46.24   447.24  1577.24
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3412.76      89.06  38.321  < 2e-16 ***
## raceBlack    -425.06     146.37  -2.904 0.004139 **
## raceOther    -409.26     111.35  -3.676 0.000312 ***
## smokeYes     -386.20     104.28  -3.704 0.000281 ***
## htYes        -472.33     197.46  -2.392 0.017768 *
## uiYes        -563.09     135.82  -4.146 5.17e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 655.4 on 183 degrees of freedom
## Multiple R-squared:  0.2136,Adjusted R-squared:  0.1922
## F-statistic: 9.944 on 5 and 183 DF,  p-value: 1.98e-08
```

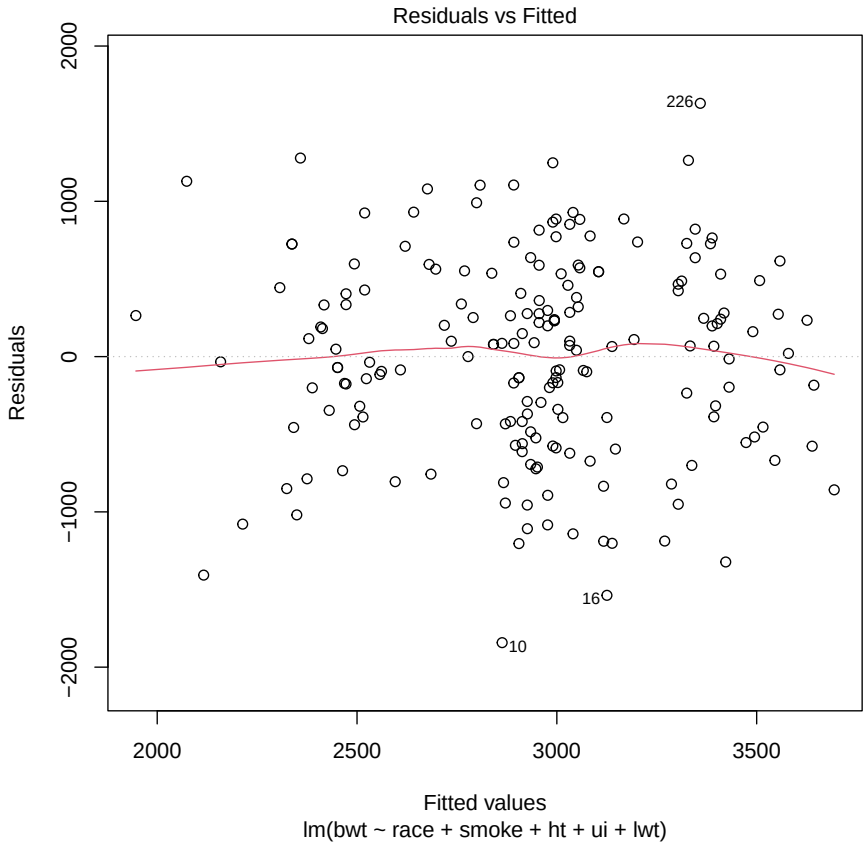
E' possibile notare un errore residuo è più alto (655.4), inoltre questo modello ha solamente variabili esplicative categoriche, pur dovendo fare regressione di una quantità continua.

Test del rapporto di verosimiglianza Per verificare se i modelli ridotti siano modelli annidati, è possibile utilizzare il test del rapporto di verosimiglianza. Per il calcolo è stato utilizzata la funzione lrtest della libreria lmtest.

```
lrtest(mq2,mq3)

## Likelihood ratio test
##
## Model 1: bwt ~ age + lwt + race + smoke + ht + ui + ptl.f + ftv.f
## Model 2: bwt ~ race + smoke + ht + ui + lwt
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  11 -1486.2
## 2   8 -1487.6 -3 2.7762     0.4274

lrtest(mq2,mq4)

## Likelihood ratio test
##
## Model 1: bwt ~ age + lwt + race + smoke + ht + ui + ptl.f + ftv.f
## Model 2: bwt ~ race + smoke + ht + ui
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  11 -1486.2
## 2   7 -1490.8 -4 9.3176    0.05363 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

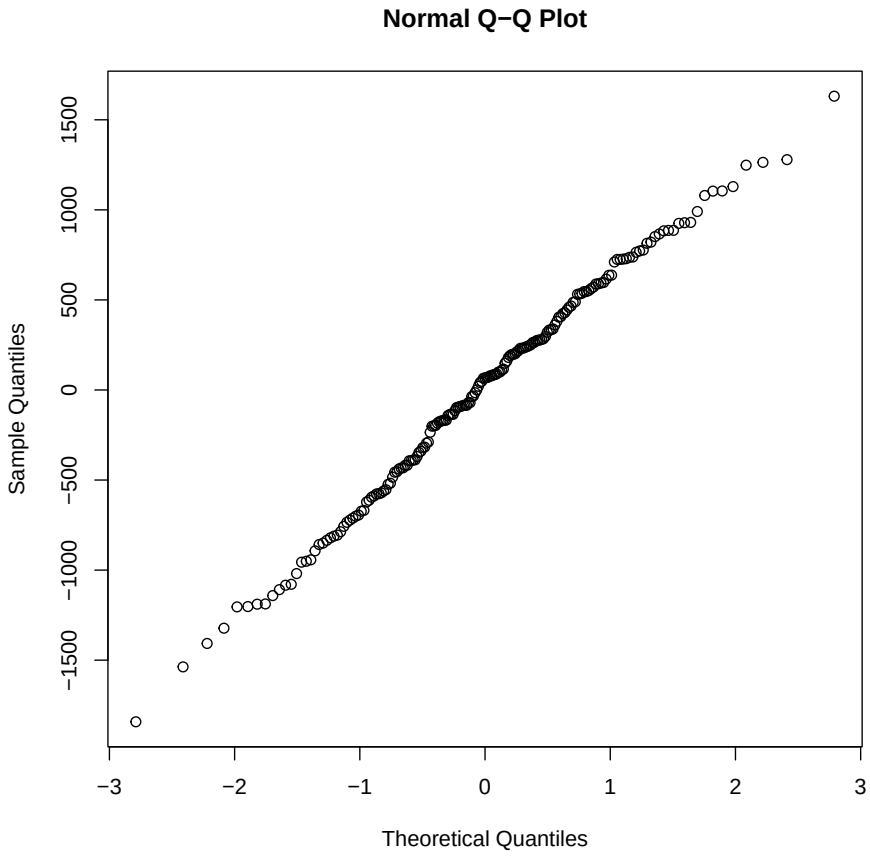
Nel modello che comprende la variabile lwt, si ha un p-value di 0.45, molto al di sopra della soglia di significatività $\alpha = 0.05$, quindi non si ha evidenza contro l'ipotesi nulla di modello annidato. Nel secondo modello, invece si ha un p-value pari a 0.056, appena sopra la soglia di significatività. Per questo motivo è stato preferito il primo modello, in quanto potrebbe garantire una migliore previsione, evitando una perdita di informazioni significative.

Analisi dei residui Possiamo analizzare i residui del modello ridotto, in modo da verificare se questi seguono un andamento lineare o mostrano un pattern differente



Il grafico rappresentante l'andamento dei residui mostra una quasi perfetta linearità. Questo fa pensare che un modello lineare riesca ad adattare bene i dati
Se al contrario avessimo visto un diverso andamento, potrebbe essere utile utilizzare un modello nel quale compaiono funzioni delle variabili in modo da ridurre l'errore.

Valutiamo l'ipotesi di distribuzione normale degli errori: allo scopo si può utilizzare il qqplot che confronta la distribuzione empirica dei residui con i quantili della distribuzione normale.



L'andamento dei punti è ben sovrapponibile con la bisettrice del grafico,rappresentante i punti della distribuzione normale, quindi l'ipotesi di distribuzione normale degli errori è confermata.

3.3 Modelli con interazioni

I modelli finora utilizzati non supponevano nessuna interazione tra le variabili. E' possibile verificare se modelli con interazioni producano una stima migliore: dato il gran numero di variabili a disposizione, è conveniente sfruttare le informazioni ottenute con l'analisi della correlazione effettuata precedentemente, in modo da analizzare solamente i casi nei quali era stata evidenziata una interazione evidente tra due variabili. Non verranno considerate interazioni con più di due variabili.

```
mq5 <- lm(bwt ~  + race * smoke + ht + lwt + ui ,data=birthwt)
summary(mq5)
```

```
##
## Call:
## lm(formula = bwt ~ +race * smoke + ht + lwt + ui, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1888.20  -387.43    32.25   417.36  1558.72
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2968.440    254.166  11.679  < 2e-16 ***
## raceBlack          -505.899    189.928  -2.664 0.008432 **
## raceOther          -481.930    135.222  -3.564 0.000468 ***
## smokeYes           -480.401    134.572  -3.570 0.000458 ***
## htYes              -540.253    200.523  -2.694 0.007723 **
## lwt                   3.763      1.687   2.231 0.026928 *
## uiYes              -548.416    135.774  -4.039 7.93e-05 ***
## raceBlack:smokeYes   39.295    293.767   0.134 0.893739
## raceOther:smokeYes  467.263    247.853   1.885 0.061009 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 643 on 180 degrees of freedom
## Multiple R-squared:  0.2556,Adjusted R-squared:  0.2225
## F-statistic: 7.724 on 8 and 180 DF,  p-value: 6.967e-09
```

```
mq6 <- lm(bwt ~  + race + smoke + ht * lwt + ui ,data=birthwt)
summary(mq6)
```

```
##
## Call:
## lm(formula = bwt ~ +race + smoke + ht * lwt + ui, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1841.99  -442.96    61.86   445.58  1612.02
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2999.144    259.446  11.560  < 2e-16 ***
## raceBlack    -489.749    145.034  -3.377 0.000898 ***
## raceOther    -349.621    111.737  -3.129 0.002045 **
## smokeYes     -379.650    103.730  -3.660 0.000331 ***
## htYes       -1791.370    718.757  -2.492 0.013590 *
## lwt             3.080      1.794   1.717 0.087713 .
## uiYes        -536.557    134.073  -4.002 9.15e-05 ***
```

```
## htYes:lwt       7.880      4.513   1.746 0.082496 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 642.3 on 181 degrees of freedom
## Multiple R-squared:  0.253,Adjusted R-squared:  0.2241
## F-statistic: 8.756 on 7 and 181 DF,  p-value: 2.955e-09
```

```
mq7 <- lm(bwt ~  + race + smoke + ht + lwt + ui * ptl.f
          ,data=birthwt)
summary(mq7)
```

```
##
## Call:
## lm(formula = bwt ~ +race + smoke + ht + lwt + ui * ptl.f, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1803.86  -427.79    37.63   484.40  1610.74
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2908.306    246.064  11.819  < 2e-16 ***
## raceBlack      -447.544    146.094  -3.063 0.002525 **
## raceOther      -333.295    112.258  -2.969 0.003395 **
## smokeYes       -327.786    104.969  -3.123 0.002089 **
## htYes          -566.564    199.044  -2.846 0.004935 **
## lwt               3.829      1.685   2.272 0.024265 *
## uiYes          -581.198    160.008  -3.632 0.000366 ***
## ptl.fYes       -290.364    154.295  -1.882 0.061467 .
## uiYes:ptl.fYes  324.680    304.480   1.066 0.287698
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 643.2 on 180 degrees of freedom
## Multiple R-squared:  0.2551,Adjusted R-squared:  0.222
## F-statistic: 7.707 on 8 and 180 DF,  p-value: 7.294e-09
```

Commenti

- I modelli hanno una stima della deviazione standard inferiore rispetto al modello senza interazioni e indice R^2 migliore, tuttavia la differenza è minima.
- In entrambi i casi le interazioni non sono risultate significative, e anche le variabili che compongono le interazioni non risultano più significative

Per questi motivi è stato preferito il modello di regressione lineare che non considera interazioni, in quanto pur essendo più semplice riesce ad adattare bene i dati.

4 Modello di regressione logistica

Invece di effettuare una regressione sul peso del bambino bwt, supponiamo di voler solamente predire se il bambino sarà sottopeso (peso inferiore a 2500 grammi), ovvero calcolare il valore atteso

$$E(\text{low}_i|\{X_i = x_i\}) = \pi_i$$

Per fare questo, si può utilizzare un modello di regressione logistica. Studiamo innanzitutto dei modelli nei quali compare una singola variabile

```
fit1 <- glm(low ~ race, family = binomial, data = birthwt)
summary(fit1)

##
## Call:
## glm(formula = low ~ race, family = binomial, data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0489  -0.9665  -0.7401   1.4042   1.6905
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.1550     0.2391  -4.830 1.36e-06 ***
## raceBlack     0.8448     0.4634   1.823   0.0683 .
## raceOther     0.6362     0.3478   1.829   0.0674 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 229.66  on 186  degrees of freedom
## AIC: 235.66
##
## Number of Fisher Scoring iterations: 4

fit2 <- glm(low ~ ht, family = binomial, data = birthwt)
summary(fit2)

##
```

```
## Call:
## glm(formula = low ~ ht, family = binomial, data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3232  -0.8341  -0.8341   1.5652   1.5652
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.8771     0.1650  -5.315 1.07e-07 ***
## htYes         1.2135     0.6083   1.995   0.0461 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 230.65  on 187  degrees of freedom
## AIC: 234.65
##
## Number of Fisher Scoring iterations: 4
```

```
fit3 <- glm(low ~ smoke, family = binomial, data = birthwt)
summary(fit3)
```

```
##
## Call:
## glm(formula = low ~ smoke, family = binomial, data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0197  -0.7623  -0.7623   1.3438   1.6599
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.0871     0.2147  -5.062 4.14e-07 ***
## smokeYes      0.7041     0.3196   2.203   0.0276 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 229.80  on 187  degrees of freedom
## AIC: 233.8
##
```

```
## Number of Fisher Scoring iterations: 4

fit4 <- glm(low ~ ui, family = binomial, data = birthwt)
summary(fit4)

##
## Call:
## glm(formula = low ~ ui, family = binomial, data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1774  -0.8097  -0.8097   1.1774   1.5967
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.9469     0.1756  -5.392 6.97e-08 ***
## uiYes         0.9469     0.4168   2.272   0.0231 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 229.60  on 187  degrees of freedom
## AIC: 233.6
##
## Number of Fisher Scoring iterations: 4

fit5 <- glm(low ~ lwt, family = binomial, data = birthwt)
summary(fit5)

##
## Call:
## glm(formula = low ~ lwt, family = binomial, data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0951  -0.9022  -0.8018   1.3609   1.9821
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.99831    0.78529   1.271   0.2036
## lwt         -0.01406    0.00617  -2.279   0.0227 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 228.69  on 187  degrees of freedom
## AIC: 232.69
##
## Number of Fisher Scoring iterations: 4

fit6 <- glm(low ~ ptl.f, family = binomial, data = birthwt)
summary(fit6)

##
## Call:
## glm(formula = low ~ ptl.f, family = binomial, data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3537  -0.7723  -0.7723   1.0108   1.6464
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.0571     0.1813  -5.831  5.5e-09 ***
## ptl.fYes      1.4626     0.4144   3.529 0.000417 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 221.90  on 187  degrees of freedom
## AIC: 225.9
##
## Number of Fisher Scoring iterations: 4

confint(fit6)

## Waiting for profiling to be done...

##                   2.5 %     97.5 %
## (Intercept) -1.4236700 -0.7110199
## ptl.fYes     0.6607048  2.2975845
```

Commenti

- Per tutte le variabili binarie e categoriche l'intercetta è negativa, mentre le variabili hanno effetto positivo, ovvero tendono a far aumentare la

probabilità che la madre abbia un bambino sottopeso.

- La variabile lwt (peso della madre) ha un effetto negativo: all'aumentare del peso della madre, la probabilità di avere un bambino sottopeso diminuisce. Questa variabile ha una sufficiente significatività.
- La variabile race non è molto significativa, con p-value 0.067, maggiore della soglia di significatività 0.05. Anche le variabili age e ftv, non riportate, non sono risultate significative.
- Le variabili ht,smoke e ui risultano significative
- La variabile ptl risulta molto significativa. In questo caso, è stata utilizzata la variabile dicotomizzata come variabile binaria: Potrebbe esserci una relazione tra le madri che hanno avuto parti prematuri e la probabilità di avere bambini sottopeso.

Possiamo inoltre calcolare un indice pseudo R^2 per il modello con la variabile ptl

$$\text{pseudo} - R^2 = 1 - \frac{\text{deviance}}{\text{null.deviance}}$$

```
pseudoRfit6 <- ((fit6$null.deviance/-2) - (fit6$deviance /-2)) / (fit6$null.deviance/-2)
pseudoRfit6

## [1] 0.05443445
```

Questo indice risulta molto basso.
E' possibile provare modelli con un maggior numero di variabili per verificare se producono risultati migliori

```
##
## Call:
## glm(formula = low ~ ht + ptl.f + lwt + ui + smoke + race + age +
##     ftv.f, family = binomial, data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6723  -0.8077  -0.5146   0.9498   2.1783
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.411346   1.894722   0.745  0.45634
## htYes        1.826891   0.705423   2.590  0.00960 **
## ptl.fYes     1.233034   0.465573   2.648  0.00809 **
## lwt         -0.014936   0.007051  -2.118  0.03414 *
## uiYes        0.705020   0.464322   1.518  0.12892
```

```
## smokeYes     0.815467   0.420198   1.941  0.05230 .
## raceBlack    1.203890   0.534233   2.253  0.02423 *
## raceOther    0.775619   0.459366   1.688  0.09132 .
## age         -0.326252   0.371046  -0.879  0.37925
## ftv.fYes    -0.125163   0.375513  -0.333  0.73890
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 196.81  on 179  degrees of freedom
## AIC: 216.81
##
## Number of Fisher Scoring iterations: 4
## [1] 0.1613434
##
## Call:
## glm(formula = low ~ ht + ptl.f + lwt + smoke + race, family = binomial,
##     data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8188  -0.8035  -0.5457   0.9667   2.1530
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.09462    0.95704   0.099  0.92124
## htYes        1.76744    0.70841   2.495  0.01260 *
## ptl.fYes     1.23144    0.44625   2.760  0.00579 **
## lwt         -0.01673    0.00695  -2.407  0.01608 *
## smokeYes     0.87611    0.40071   2.186  0.02879 *
## raceBlack    1.26372    0.52933   2.387  0.01697 *
## raceOther    0.86418    0.43509   1.986  0.04701 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 200.48  on 182  degrees of freedom
## AIC: 214.48
##
## Number of Fisher Scoring iterations: 4
##
```

```
## Call:
## glm(formula = low ~ ht + ptl.f + lwt, family = binomial, data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7420  -0.8018  -0.6895   0.9647   2.2460
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.017367   0.853337   1.192  0.23317
## htYes        1.893971   0.721090   2.627  0.00863 **
## ptl.fYes     1.406770   0.428501   3.283  0.00103 **
## lwt         -0.017280   0.006787  -2.546  0.01090 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 210.12  on 185  degrees of freedom
## AIC: 218.12
##
## Number of Fisher Scoring iterations: 4
##
## Call:
## glm(formula = low ~ ht + ptl.f, family = binomial, data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3230  -0.7398  -0.7398   1.0385   1.6909
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.1560     0.1911  -6.048 1.47e-09 ***
## htYes         1.2879     0.6269   2.054 0.039940 *
## ptl.fYes      1.4919     0.4193   3.558 0.000373 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 217.66  on 186  degrees of freedom
## AIC: 223.66
##
```

```
## Number of Fisher Scoring iterations: 4
```

Commenti

- Nel modello completo le variabili ui, age , ftv, race e smoke non sono significative. L'indice pseudo-R^2 è rimasto relativamente basso
- Togliendo le variabili ui, age e ftv tutte le restanti variabili diventano significative. L'effetto delle variabili mantiene lo stesso segno trovato in precedenza.
- Escludendo le variabili meno significative, sono stati trovati modelli sempre più ridotti, nei quali tutte le variabili hanno una ottima significatività.

Per aiutare a capire se i modelli ridotti siano utilizzabili, si possono nuovamente effettuare dei test di rapporto di verosimiglianza

```
lrtest(fit7,fit8,fit9,fit10,fit6)

## Likelihood ratio test
##
## Model 1: low ~ ht + ptl.f + lwt + ui + smoke + race + age + ftv.f
## Model 2: low ~ ht + ptl.f + lwt + smoke + race
## Model 3: low ~ ht + ptl.f + lwt
## Model 4: low ~ ht + ptl.f
## Model 5: low ~ ptl.f
##   #Df   LogLik Df  Chisq Pr(>Chisq)
## 1  10  -98.405
## 2   7 -100.241 -3 3.6730   0.299003
## 3   4 -105.062 -3 9.6411   0.021876 *
## 4   3 -108.831 -1 7.5386   0.006039 **
## 5   2 -110.949 -1 4.2358   0.039579 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Il test del rapporto di verosimiglianza tra il modello completo e quello che esclude le variabili ui,age e ftv non fornisce evidenza contro il modello ridotto.
- Togliendo ulteriori variabili il test porterebbe a rifiutare i modelli ridotti
- Questa scelta dovrà essere approfondita e validata utilizzando algoritmi stepwise e attraverso modelli grafici

Calcoliamo l'indice pseudo-R^2 per i modelli ridotti low $\sim$ ht + ptl.f + lwt

```
pseudoRfit9 <- ((fit9$null.deviance/-2) - (fit9$deviance /-2)) / (fit9$null.deviance/-2)
pseudoRfit9

## [1] 0.1046082
```

e low $\sim$ ht + ptl + lwt + smoke + race

```
pseudoRfit8 <- ((fit8$null.deviance/-2) - (fit8$deviance /-2)) / (fit8$null.deviance/-2)
pseudoRfit8

## [1] 0.1456916
```

Pur rimanendo abbastanza basso, l'indice è notevolmente migliore rispetto al modello di regressione logistica con la sola variabile ptl, e più alto rispetto al modello low $\sim$ ht + ptl.f + lwt Possiamo valutare un intervallo di confidenza dei parametri

```
confint(fit8)

## Waiting for profiling to be done...

##                    2.5 %       97.5 %
## (Intercept) -1.73046202  2.037044326
## htYes        0.41461419  3.246747983
## ptl.fYes     0.36517972  2.126748119
## lwt         -0.03123596 -0.003854403
## smokeYes     0.10090714  1.681646805
## raceBlack    0.22732716  2.319514275
## raceOther    0.02399401  1.739913400
```

Tutte le variabili hanno un intervallo di confidenza che non include lo zero.

Stima della probabilità Possiamo stimare la probabilità $\hat{\pi}_i$ di una madre di avere un figlio sottopeso: Prima vediamo la probabilità per una madre bianca che non ha avuto parti prematuri, senza familiarità di ipertensione, e con un peso vicino alla media (60 kg)

```
stima1 <- exp(coef(fit8)%*%c(1,0,0,132.3,0,0,0))/
  (1+exp(coef(fit8)%*%c(1,0,0,132.3,0,0,0)))
stima1

##           [,1]
## [1,] 0.1073036
```

Vediamo adesso la stima per una madre di colore, che ha avuto parti prematuri, con una familiarità di ipertensione e peso vicino alla media (60 kg)

```
stima2 <- exp(coef(fit8)%*%c(1,1,1,132.3,1,1,0))/
  (1+exp(coef(fit8)%*%c(1,1,1,132.3,1,1,0)))
stima2

##            [,1]
## [1,] 0.9534751
```

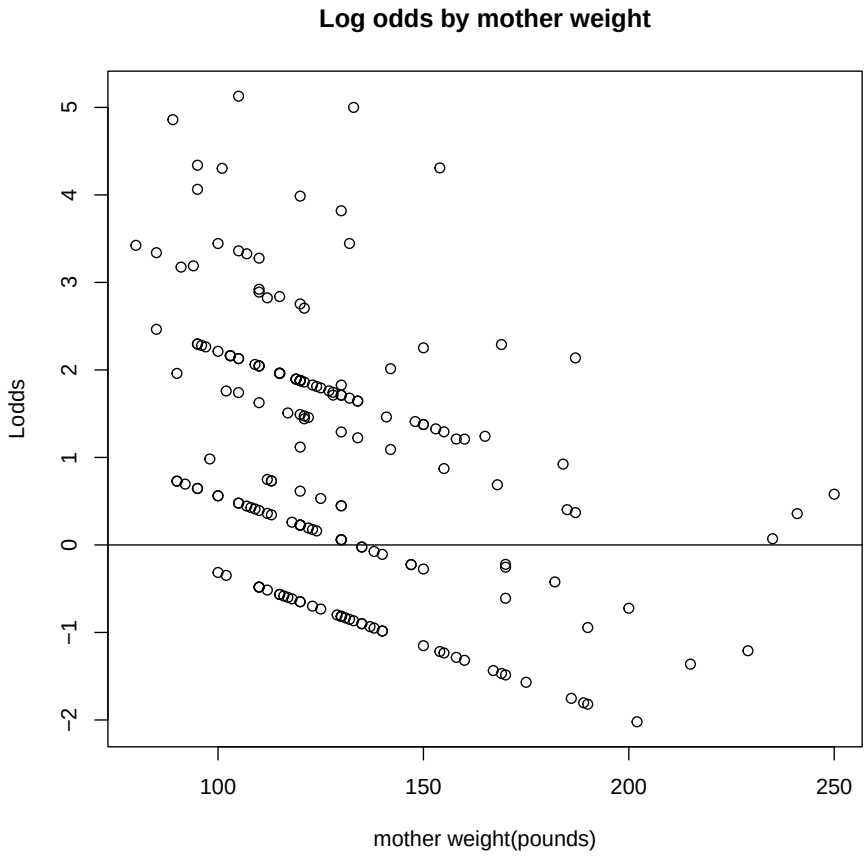
La probabilità nel secondo caso è del 95%.

Odds Possiamo calcolare una stima degli odds nel secondo caso

$$\text{odds} = \frac{\pi}{1-\pi} = \frac{0.953}{0.047} = 20,28$$

E' possibile vedere l'andamento del logaritmo degli odds all'aumentare del peso della madre

```
coeff <- fit8$coefficients
Lodds <- coeff[1] + coeff[2]*src$ht + coeff[3]*ptl.c + coeff[4]*src$lwt + coeff[5]*src$smoke

plot(src$lwt,Lodds,xlab="mother weight(pounds)",main="Log odds by mother weight")
abline(0,0)
```



Pur notando un andamento decrescente, i punti risultano molto sparsi. Si nota anche che il numero di campioni diminuisce all'aumentare del peso, cosa che rende più difficile verificare la relazione tra gli odds e il peso della madre

4.1 Modelli con interazioni

Verifichiamo se sono presenti interazioni tra le variabili, utilizzando le variabili migliori individuate precedentemente. La correlazione studiata in precedenza non ha individuato tuttavia forti legami tra le variabili

```
fit11 <- glm(low ~ ht + ptl.f + lwt + smoke * race
             , family = binomial
             , data = birthwt)
summary(fit11)

##
```

```
## Call:
## glm(formula = low ~ ht + ptl.f + lwt + smoke * race, family = binomial,
##     data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9131  -0.8316  -0.4734   0.9316   2.2737
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.260784   1.066427  -0.245  0.80681
## htYes               1.705723   0.720469   2.368  0.01791 *
## ptl.fYes            1.247869   0.450046   2.773  0.00556 **
## lwt                -0.016273   0.007115  -2.287  0.02219 *
## smokeYes            1.285878   0.623711   2.062  0.03924 *
## raceBlack           1.423816   0.785230   1.813  0.06979 .
## raceOther           1.313087   0.620635   2.116  0.03437 *
## smokeYes:raceBlack -0.072691   1.097794  -0.066  0.94721
## smokeYes:raceOther -1.179176   0.929137  -1.269  0.20440
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 198.64  on 180  degrees of freedom
## AIC: 216.64
##
## Number of Fisher Scoring iterations: 5
```

```
fit12 <- glm(low ~ ht * lwt + ptl.f  + smoke + race
             , family = binomial
             , data = birthwt)
summary(fit12)
```

```
##
## Call:
## glm(formula = low ~ ht * lwt + ptl.f + smoke + race, family = binomial,
##     data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8197  -0.7907  -0.5443   0.9525   2.1519
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  0.048737   1.048168   0.046   0.9629
## htYes        2.038451   2.670620   0.763   0.4453
## lwt         -0.016378   0.007673  -2.134   0.0328 *
## ptl.fYes     1.229462   0.446557   2.753   0.0059 **
## smokeYes     0.881596   0.403808   2.183   0.0290 *
## raceBlack    1.265805   0.529435   2.391   0.0168 *
## raceOther    0.865844   0.435183   1.990   0.0466 *
## htYes:lwt   -0.001788   0.016979  -0.105   0.9161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 200.47  on 181  degrees of freedom
## AIC: 216.47
##
## Number of Fisher Scoring iterations: 4
```

```
fit13 <- glm(low ~ ht + lwt + ptl.f + age * ftv.c
             , family = binomial
             , data = birthwt)
summary(fit13)
```

```
##
## Call:
## glm(formula = low ~ ht + lwt + ptl.f + age * ftv.c, family = binomial,
##     data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7528  -0.7782  -0.5893   0.8921   2.3602
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.96341    2.13666  -0.451 0.652065
## htYes        1.81523    0.73274   2.477 0.013238 *
## lwt         -0.01823    0.00729  -2.500 0.012407 *
## ptl.fYes     1.60258    0.45686   3.508 0.000452 ***
## age          0.46338    0.45907   1.009 0.312787
## ftv.c        5.96442    2.03407   2.932 0.003365 **
## age:ftv.c   -1.28249    0.44033  -2.913 0.003585 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 196.30  on 182  degrees of freedom
## AIC: 210.3
##
## Number of Fisher Scoring iterations: 5
```

L'unica interazione che è stata provata come significativa è quella tra età e numero di visite effettuate dal ginecologo nel primo trimestre (ftv). Tuttavia, possiamo vedere come l'età da sola non sia una variabile significativa, inoltre le variabili age e ftv erano state scartate nelle precedenti analisi. Come nella regressione lineare, la scelta è orientata verso il modello privo di interazioni.

5 Selezione del modello

Per confermare quanto ottenuto effettuando una selezione informata dei modelli, si possono utilizzare degli algoritmi stepwise sfruttando diversi criteri di selezione del modello. Saranno impiegati i criteri di selezione AIC (*Akaike's information criterion*) e BIC (*Bayesian information criterion*),e sarà effettuata una ricerca del modello in avanti (*forward*) partendo dal modello nullo con la sola intercetta, all'indietro (*backward*) partendo dal modello completo e mista (*both*), partendo dal modello nullo ma con possibilità di rimuovere variabili aggiunte durante le iterazioni.

5.1 Modelli di regressione lineare

Troviamo un modello di regressione della variabile bwt, utilizzando le variabili ptl e ftv come ordinali, age e lwt come interi ed escludendo dal dataset la variabile low

```
reg.data = birthwt[c(2,3,4,5,6,7,8,9,10)]
mq0 <- lm(bwt ~ 1,data=reg.data)
mq.sat <- lm (bwt ~age+lwt+race+smoke+ptl+ht+ui+ftv,data=reg.data)
```

5.1.1 Metodo Forward

Partendo dal modello nullo con la sola intercetta, effettuiamo una procedura stepwise per la selezione del modello.

```
forw_aic <- step(mq0,scope = formula(mq.sat)
                 , direction="forward", k=2)

forw_bic <- step(mq0,scope = formula(mq.sat)
                 , direction="forward", k=log(length(reg.data$bwt)))
```

Il metodo AIC termina con il modello

$$\text{bwt} \sim \text{ui} + \text{race} + \text{smoke} + \text{ht} + \text{lwt} + \text{ptl}$$

Il modello BIC termina con il modello

$$\text{bwt} \sim \text{ui} + \text{race} + \text{smoke} + \text{ht} + \text{lwt}$$

Questo secondo modello è lo stesso che era stato scelto durante l'analisi in 3.2, fornendo ulteriore conferma della bontà del modello scelto.
Il modello selezionato con criterio AIC e quello selezionato con criterio BIC differiscono per la variabile ptl, non selezionata secondo il criterio BIC, solitamente più parsimonioso.
Effettuando una ricerca mista che permette la rimozione di variabili, si trovano gli stessi modelli

5.1.2 Metodo Backward

Partendo dal modello saturo, si effettua una procedura stepwise per la selezione del modello.

```
back_aic <- step(mq.sat,scope = formula(mq.sat)
                 , direction="backward", k=2, trace=FALSE)
back_bic <- step(mq.sat,scope = formula(mq.sat)
                 , direction="backward", k=log(length(reg.data$bwt)), trace=FALSE)
```

In entrambi i casi, vengono trovati gli stessi modelli individuati dal metodo Forward.

5.2 Modelli di regressione logistica

Troviamo adesso un modello di regressione logistica per la variabile low, utilizzando le stesse variabili utilizzate nella sezione precedente ed escludendo dal dataset la variabile bwt.

```
class.data = birthwt[c(1,2,3,4,5,6,7,8,9)]

cq0 <- glm(unclass(low) ~ 1
           ,data=class.data)
cq.sat <- glm (unclass(low) ~ age+lwt+race+smoke+ptl.f+ht+ui+ftv
               ,data=class.data)
```

5.2.1 Metodo Forward

Partendo dal modello con la sola intercetta, si effettua una procedura stepwise in avanti

```
class_forw_aic <- step(cq0,scope = formula(cq.sat)
                       , direction="forward", k=2)
class_forw_bic <- step(cq0,scope = formula(cq.sat)
                       , direction="forward", k=log(length(class.data$low)))
```

Il criterio AIC termina con il modello

$$\text{low} \sim \text{ptl.f} + \text{ht} + \text{lwt} + \text{ui} + \text{race} + \text{smoke}$$

Questo modello è molto simile a quello scelto in 4, tuttavia in questo caso viene inclusa anche la variabile ui che era invece stata scartata durante l'analisi.

Mentre il criterio BIC termina con il modello

$$\text{low} \sim \text{ptl.f}$$

In 4 era stato notato che la variabile ptl fosse particolarmente significativa. Tuttavia, era anche dimostrato che un modello con quest'unica variabile binaria avesse un basso indice pseudo-R^2,e che non fosse un modello annidato rispetto al modello completo.

5.2.2 Metodo backward

```
class_back_aic <- step(cq.sat,scope = formula(cq.sat)
                       , direction="backward", k=2)
class_back_bic <- step(cq.sat,scope = formula(cq.sat)
                       , direction="backward", k=log(length(reg.data$bwt)))
```

Il criterio AIC termina con lo stesso modello visto nella direzione forward. Il criterio BIC termina invece con un modello diverso:

$$\text{low} \sim \text{lwt} + \text{ptl.f} + \text{ht}$$

Viene riproposto un modello già trovato in 4. Anche questo modello non era risultato annidato per il test del rapporto di verosimiglianza, ed era stato preferito un modello più complesso.

6 Modelli grafici

Utilizzando i modelli grafici è possibile valutare l'indipendenza condizionata tra le variabili e fornire una rappresentazione efficace del modello statistico. I modelli grafici verranno utilizzati per trovare le dipendenze tra la variabile low (bambino sottopeso) e le restanti variabili. Per fare questo verranno utilizzati modelli grafici per variabili categoriche.

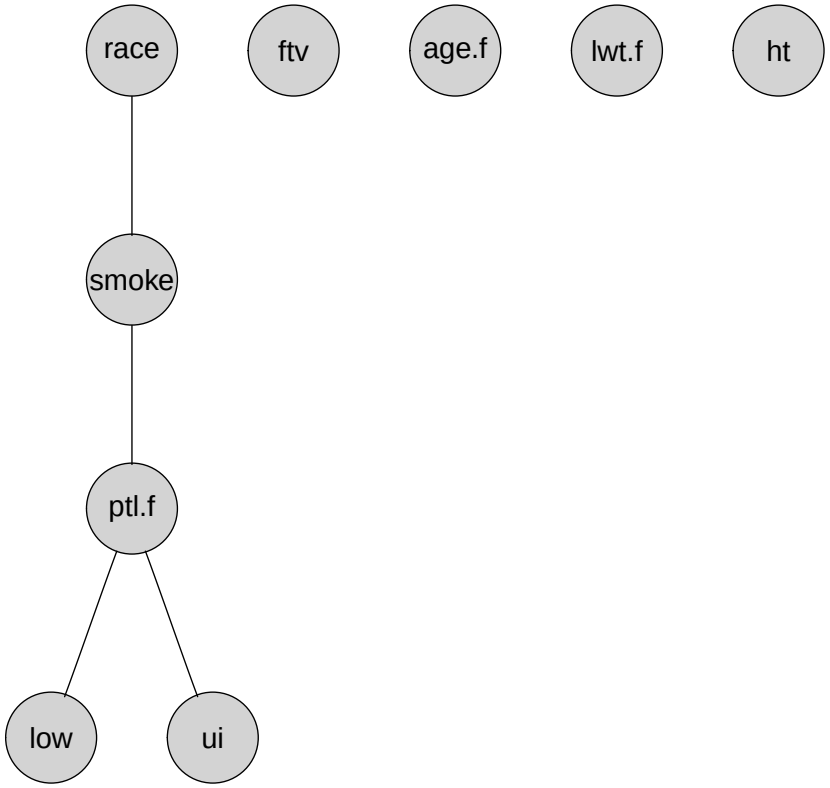
6.1 Undirected Graphs

Per trovare le relazioni di indipendenza condizionata tra le variabili (categoriche) e la variabile obiettivo low, sono state sfruttate le seguenti variabili dicotomizzate

- age, dicotomizzata in Young (Età compresa tra 14 e 23 anni), Adult (Età compresa tra 24 e 35 anni) e Over 35.
- ptl, dicotomizzata in Yes (La madre ha avuto precedenti parti prematuri) e No altrimenti.
- ftv, dicotomizzata in Yes (La madre ha effettuato visite ginecologiche nel primo trimestre) e No altrimenti

La scelta di dicotomizzare ptl e ftv come variabili binarie e non come variabili ordinali si è rivelata utile alla stabilità del metodo, poichè senza questa dicotomizzazione,i modelli trovati mostravano una completa indipendenza tra le variabili esplicative e la variabile low. Era stato notato in precedenza come la variabile ptl fosse maggiormente significativa se dicotomizzata in variabile binaria.

Non effettuando ipotesi a priori sulla struttura del grafo, è possibile far apprendere la struttura del grafo dai dati attraverso una procedura iterativa. E' possibile sfruttare i criteri AIC e BIC visti precedentemente per la penalizzazione della verosimiglianza.
Il modello calcolato con BIC è il seguente



Questo modello mostra una completa indipendenza delle variabili ftv, age, lwt e ht. E' possibile verificare l'ipotesi di indipendenza marginale di tutte le variabili dalla variabile obiettivo,low attraverso dei test di ipotesi

```
ciTest(data,set=~low + ht)

## Testing low _|_ ht
## Statistic (DEV):    4.022 df: 1 p-value: 0.0449 method: CHISQ

ciTest(data,set=~low + age.f)

## Testing low _|_ age.f
## Statistic (DEV):    3.805 df: 2 p-value: 0.1492 method: CHISQ

ciTest(data,set=~low + lwt.f)

## Testing low _|_ lwt.f
## Statistic (DEV):    2.508 df: 1 p-value: 0.1133 method: CHISQ
```

```
ciTest(data,set=~low + ftv)

## Testing low _|_ ftv
## Statistic (DEV):    6.186 df: 5 p-value: 0.2886 method: CHISQ

ciTest(data,set=~low + race)

## Testing low _|_ race
## Statistic (DEV):    5.010 df: 2 p-value: 0.0817 method: CHISQ

ciTest(data,set=~low + ui)

## Testing low _|_ ui
## Statistic (DEV):    5.076 df: 1 p-value: 0.0243 method: CHISQ

ciTest(data,set=~low + ptl.f)

## Testing low _|_ ptl.f
## Statistic (DEV):   12.774 df: 1 p-value: 0.0004 method: CHISQ

ciTest(data,set=~low + smoke)

## Testing low _|_ smoke
## Statistic (DEV):    4.867 df: 1 p-value: 0.0274 method: CHISQ
```

- I test di indipendenza condizionale non forniscono evidenza contro l'ipotesi nulla di indipendenza marginale di race, age, lwt e ftv da low. Possiamo notare non sia marginalmente indipendente da low.
- I test di indipendenza condizionale forniscono evidenza contro l'ipotesi nulla di indipendenza marginale di ui,smoke,ht e ptl da low, a conferma dei collegamenti presenti nel grafo
- La variabile ht è completamente indipendente nel grafo, ma tale ipotesi non è confermata dal test.
  Notiamo tuttavia che il p-value è molto vicino alla soglia di significatività $\alpha = 0.05$
- Il modello di regressione logistica e low $\sim$ ht + ptl + lwt + smoke + race includeva le variabili ht e lwt, che in questo caso risultano marginalmente indipendenti da low.
  Tuttavia viene confermata la forte relazione tra ptl e la variabile obiettivo low e viene spiegato il motivo della selezione di modelli con solamente la variabile ptl, poichè è l'unica che condiziona low.

Questa struttura inoltre fa sì che race sia condizionalmente indipendente da low dato smoke e ptl, e smoke e ui siano condizionalmente indipendenti da low dato ptl.

```
low.ug <- ug(~smoke*race + ptl.f*smoke + ui*ptl.f + low*ptl.f)
separates("race","low",c("ptl.f","smoke"),low.ug)

## [1] TRUE

separates("smoke","low",c("ptl.f"),low.ug)

## [1] TRUE

separates("ui","low",c("ptl.f"),low.ug)

## [1] TRUE
```

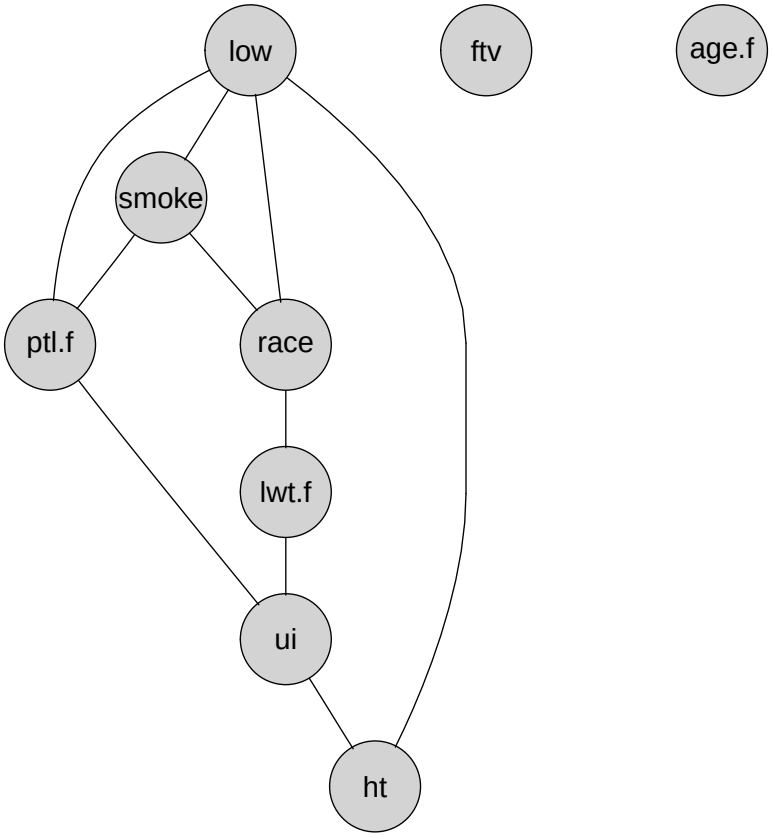
Si valuta inoltre l'ipotesi di indipendenza condizionale di smoke e ui:

```
ciTest(data,set=~low + ui +ptl.f)

## Testing low _|_ ui | ptl.f
## Statistic (DEV):    4.823 df: 2 p-value: 0.0897 method: CHISQ
## Slice information:
##   statistic p.value df ptl.f
## 1    4.7177 0.02985  1    No
## 2    0.1052 0.74568  1   Yes

ciTest(data,set=~low + smoke +ptl.f)

## Testing low _|_ smoke | ptl.f
## Statistic (DEV):    2.621 df: 2 p-value: 0.2697 method: CHISQ
## Slice information:
##   statistic p.value df ptl.f
## 1    1.7904  0.1809  1    No
## 2    0.8307  0.3621  1   Yes

ciTest(data,set=~low + race +  smoke + ptl.f)

## Testing low _|_ race | smoke ptl.f
## Statistic (DEV):   14.701 df: 8 p-value: 0.0652 method: CHISQ
## Slice information:
##   statistic p.value df smoke ptl.f
## 1    7.3630 0.02519  2    No    No
## 2    3.6936 0.15774  2   Yes    No
## 3    3.2779 0.19418  2    No   Yes
## 4    0.3669 0.83239  2   Yes   Yes
```

I test non forniscono evidenza contro le ipotesi di indipendenza.

E' possibile ripetere l'analisi con una procedura forward, partendo dal modello di completa indipendenza ed aggiungendo nuovi archi in modo iterativo.

Utilizzando il criterio BIC si ottiene lo stesso modello visto precedentemente, mentre con il criterio AIC si ottiene il seguente modello



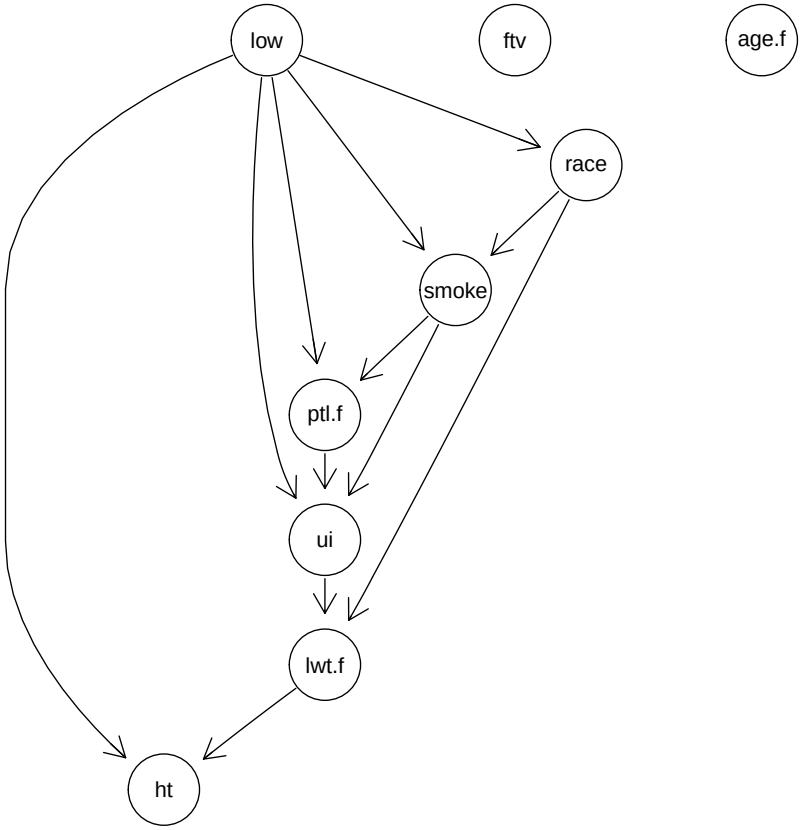
- low è condizionalmente indipendente dal peso della madre date le restanti variabili
- low è condizionalmente indipendente dall'irritabilità uterina date le restanti variabili
- Il numero di visite dal ginecologo effettuate nel primo trimestre e l'età (dicotomizzata) risultano marginalmente indipendenti

6.2 Directed Acyclic Graphs

Per studiare la propagazione delle probabilità condizionali lungo il grafo e capire come alcune evidenze modificano la probabilità della variabile obiettivo low, è possibile utilizzare un modello grafico basato su un DAG, ovvero una rete

Bayesiana.
Come effettuato in precedenza, la struttura del grafo potrà essere appresa dai dati attraverso un algoritmo hill-climbing.

Con un approccio "Naive", potremmo pensare di non impostare nesssun vincolo sugli archi e utilizzando il criterio AIC per la penalizzazione otterremmo il seguente modello

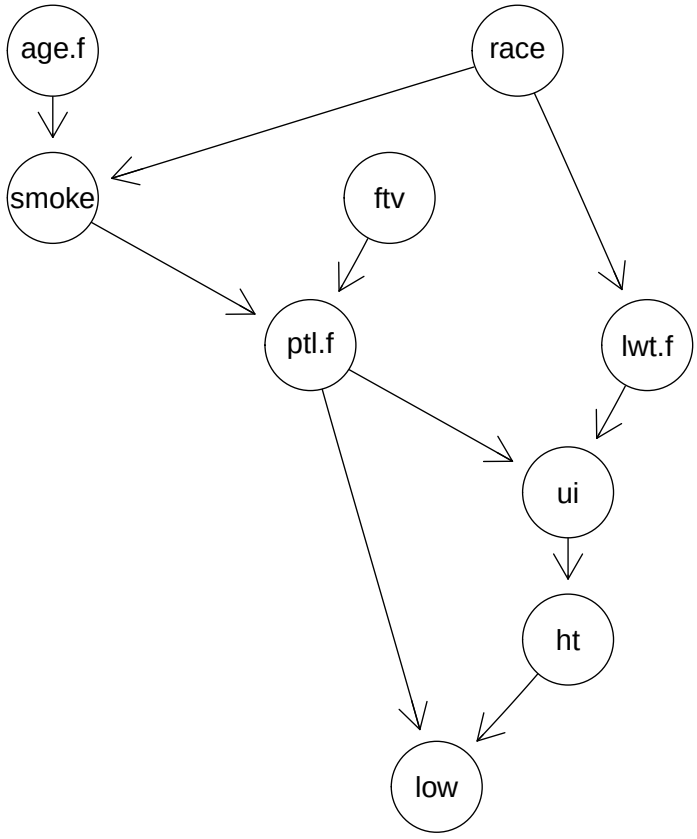


Tuttavia, è possibile notare che in questo modello si afferma che low determini l'etnia delle madri e le abitudini sul fumo. Non è possibile accettare una affermazione di questo tipo, pertanto è possibile vietare alcuni archi, in base all'ordine temporale che è necessario assumere sulle variabili.
Le variabili verranno suddivise nei seguenti gruppi

1. background: race,age,ftv,lwt
2. fattori: smoke,ht,ui,ptl
3. obiettivo:low

Non saranno permessi archi dal gruppo 2 verso il gruppo 1 e dal gruppo 3 verso 2 e 1. Lanciando nuovamente la procedura di apprendimento del grafo dai dati, si ottiene il seguente DAG

```
block <- c(3,1,2,2,2,2,1,1,1)
blM <- matrix(0,nrow=9,ncol=9)
rownames(blM) <- colnames(blM) <- names(data)
for (b in 2:3) blM[block==b,block<b] <- 1

blackL <- data.frame(get.edgelist(as(blM,"igraph")))
names(blackL) <- c("from", "to")

birthwt.bn1 <- hc(data,blacklist=blackL,score="aic")
plot(as(amat(birthwt.bn1),"graphNEL"))
```



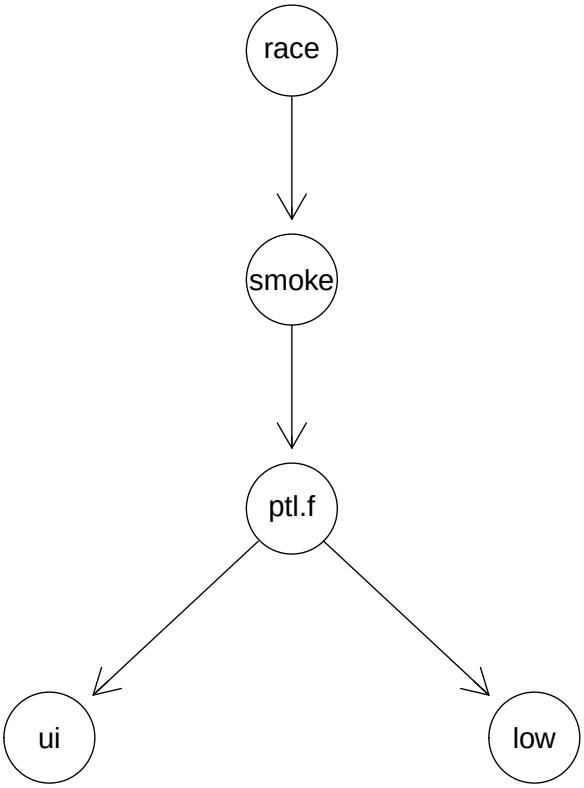
Questo modello rispetta l'ordine temporale che è stato supposto, fornendo

una rappresentazione efficace delle indipendenze condizionali.

```
dSep(amat(birthwt.bn1),"ftv","low",NULL)

## [1] FALSE

dSep(amat(birthwt.bn1),"age.f","low",NULL)

## [1] FALSE

dSep(amat(birthwt.bn1),"ftv","low",c("ptl.f","ui"))

## [1] TRUE

dSep(amat(birthwt.bn1),"age.f","low",c("ptl.f","smoke","ui"))

## [1] TRUE
```

Con l'obiettivo di ottenere un modello più semplice su cui fare inferenza, viene ripetuto lo studio utilizzando il criterio di selezione BIC

```
birthwt.bn2 <- hc(data,blacklist=blackL,score="bic")
#plot(as(amat(birthwt.bn2),"graphNEL"))
low.dag <- dag(~smoke*race + ptl.f*smoke + ui*ptl.f + low*ptl.f)
plot(low.dag)
```



Questo modello riprende la struttura degli archi vista nel modello non orientato ed è in linea con gli studi precedenti sull'indipendenza marginale delle variabili.
Utilizzando grain è possibile effettuare delle query in modo da ottenere la probabilità condizionata dall'evidenza di alcune variabili esplicative

- Probabilità marginale di avere un bambino sottopeso

```
## $low
## low
##        No       Yes
## 0.6878307 0.3121693
```

- Probabilità condizionale di avere avere un bambino sottopeso dati precedenti parti prematuri

```
##      ptl.f
## low          No Yes
##   No  0.7421384 0.4
##   Yes 0.2578616 0.6
```

- Probabilità condizionale di avere un bambino sottopeso, date le abitudini di fumo

```
##      smoke
## low         No       Yes
##   No  0.706437 0.6589155
##   Yes 0.293563 0.3410845
```

- Probabilità condizionale di avere un bambino sottopeso, data l'etnia della madre

```
##      race
## low       White     Black     Other
##   No  0.6806962 0.6881595 0.6979257
##   Yes 0.3193038 0.3118405 0.3020743
```

  Possiamo notare come le probabilità siano estremamente simili tra loro e pressochè identiche alla probabilità marginale di low.

- Probabilità condizionale di avere un bambino sottopeso, dato il fumo e la presenza di irritabilità uterina

```
## , , ui = No
##
##      smoke
## low          No       Yes
##   No  0.7131353 0.6725039
##   Yes 0.2868647 0.3274961
##
## , , ui = Yes
##
##      smoke
## low          No       Yes
##   No  0.6647129 0.5893453
##   Yes 0.3352871 0.4106547
```

Infine, è possibile provare a prevedere quali bambini saranno sottopeso sfruttando la rete appresa. I dati verranno divisi effettuando un campionamento casuale dai 189 campioni, ed utilizzandone il 75% per l'addestramento e il 25%

per la validazione, quindi il modello verrà riaddestrato solamente sul dataset di addestramento e valutato sul dataset di validazione.

Verrà quindi valutata la differenza tra il numero di casi predetti dal modello e il valore effettivo di low

```
training_size= 0.75
training_rows <- sample(seq_len(nrow(data))
                        , size = floor(training_size * nrow(data)))

train <-data[training_rows,]
val <-data[-training_rows,]
```

```
lowmod2 <- compile(grain(low.dag,data=train))

pred <- data.frame(predict(lowmod2,resp="low",newdata=val
                           ,type="class"))

table(val$low)
```

```
##
##  No Yes
##  33  15
```

```
table(val$low)/sum(table(val$low))
```

```
##
##     No    Yes
## 0.6875 0.3125
```

```
table(pred$low)
```

```
##
##  No Yes
##  39   9
```

```
tt <- table(val$low,pred$low)
tt
```

```
##
##       No Yes
##   No  31   2
##   Yes  8   7
```

```
sweep(tt,1,apply(tt,1,sum),FUN="/")
```

```
##
##               No        Yes
##   No  0.93939394 0.06060606
##   Yes 0.53333333 0.46666667
```

Il modello ha una performance mediocre: circa il 40% dei bambini sottopeso sono stati classificati correttamente. E' possibile ripetere la procedura sfruttando il DAG appreso con il criterio AIC.
Per evitare casi nei quali non siano presenti osservazioni, è stato introdotto un fattore di smoothing pari a 0.1

```
low.dag2 <- dag(~smoke*age.f + ptl.f*smoke + ui*ptl.f + low*ptl.f + ptl.f*ftv +ht*ui + ui*lw

lowmod3 <- compile(grain(low.dag2,data=train,smooth=0.1))

pred2 <- data.frame(predict(lowmod3,resp="low",newdata=val,type="class"))

table(pred2$low)

##
##  No Yes
##  47   1

tt2 <- table(val$low,pred2$low)
tt2

##
##       No Yes
##   No  33   0
##   Yes 14   1

sweep(tt2,1,apply(tt2,1,sum),FUN="/")

##
##               No        Yes
##   No  1.00000000 0.00000000
##   Yes 0.93333333 0.06666667
```

In questo caso le performance sono peggiori, quindi è preferibile il modello selezionato attraverso il criterio BIC.

7 Conclusioni

- Il numero di visite effettuate dal ginecologo durante il primo trimestre e l'età delle madri non hanno un effetto significativo sul peso dei neonati
- Una madre che ha avuto precedenti parti prematuri ha una maggiore probabilità di avere un bambino
- Un modello di regressione lineare bwt $\sim$ race + smoke + ht + ui + lwt si adatta sufficientemente bene ai dati, pur mostrando un basso indice R^2.

- Un possibile modello per la classificazione dei bambini sottopeso è il modello di regressione logistica low $\sim$ ht + ptl + lwt + smoke + race, ottenuto come miglior compromesso tra complessità e correttezza del modello.
- La rete bayesiana scelta ha confermato la relazione tra il numero di parti prematuri e la probabilità di avere un bambino sottopeso. Il peso della madre e la familiarità con l'ipertensione non compaiono nel modello più semplice, ma ottengono prestazioni comparabili ad un modello più complesso che include tutte le variabili.