Studio di regressione lineare multipla e regressione logistica multipla

Sofia Galante

1 Introduzione

Il dataset scelto come oggetto di studio è il dataset *"birthwt"* presente nel package *"MASS"*. Le variabili presenti in questo dataset sono:

1. **low** - 1 se il bambino alla nascita ha un peso troppo basso (sotto 2,5 kg), 0 altrimenti;
2. **age** - l'età della madre espressa in anni;
3. **lwt** - il peso della madre espresso in once alla fine dell'ultimo periodo mestruale;
4. **race** - 1 se la madre è di razza bianca, 2 se è di razza nera e 3 se la razza è un'altra;
5. **smoke** - 1 se la madre fuma, 0 altrimenti;
6. **ptl** - numero di precedenti parti prematuri;
7. **ht** - storia familiare di ipertensione (1 se è presente, 0 se è assente);
8. **ui** - 1 se è presente irritabilità uterina, 0 altrimenti;
9. **ftv** - numero di visite dal ginecologo nel primo trimestre;
10. **bwt** - peso alla nascita del bambino (in grammi).

Si noti inoltre che **low** è dicotomizzato da **bwt**.

```
library(BBmisc)

data("birthwt", package = 'MASS')
study <- birthwt
study$bwt <- normalize(study$bwt)
study$age <- normalize(study$age)
study$lwt <- normalize(study$lwt)
study$ftv <- normalize(study$ftv)
study$ptl <- normalize(study$ptl)
str(study)
```

```
## 'data.frame': 189 obs. of  10 variables:
##  $ low  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age  : num  -0.8 1.842 -0.611 -0.422 -0.989 ...
##  $ lwt  : num  1.707 0.824 -0.811 -0.713 -0.746 ...
##  $ race : int  2 3 1 1 1 3 1 3 1 1 ...
##  $ smoke: int  0 0 1 1 1 0 0 0 1 1 ...
##  $ ptl  : num  -0.397 -0.397 -0.397 -0.397 -0.397 ...
##  $ ht   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ui   : int  1 0 0 1 1 0 0 0 0 0 ...
##  $ ftv  : num  -0.749 2.083 0.195 1.139 -0.749 ...
##  $ bwt  : num  -0.578 -0.54 -0.532 -0.481 -0.473 ...
```

Lo studio svolto si divide in due parti: uno studio di *regressione lineare multipla* (in cui la variabile obiettivo è **bwt**) e uno studio di *regressione logistica multipla* (in cui la variabile obiettivo è **low**).
Visto che le due variabili sono una la dicotomizzazione dell'altra, si è deciso di escludere la prima dallo studio della seconda e viceversa. Inoltre, ciò che ci si aspetta è che i due modelli finali di regressione utilizzino le stesse variabili o che **low** dipenda da un sottoinsieme di variabili utilizzate nello studio di **bwt**.

1.1 Creazione dei dataset per lo studio

Il primo passo che si è svolto per lo studio del database è la creazione di due variabili dummy per la variabile **race**. Essa assume infatti 3 valori distinti (1, 2, 3) a seconda della razza di appartenenza della madre. L'idea è stata quella di creare una variabile dummy per la **white race** e una variabile dummy per la **black race**.

```
study$white <- ifelse(study$race == '1', 1, 0)
study$black <- ifelse(study$race == '2', 1, 0)
```

Una volta create e inserite le variabili dummy, si sono creati i due dataset corretti: uno per lo studio della *regressione lineare multipla* (**bwt** come variabile obiettivo) e uno per lo studio della *regressione logistica multipla* (**low** come obiettivo). Si noti che in entrambi i dataset è stata eliminata la variabile **race** e sostituita con le variabili dummy create in precedenza.

```
study <- as.data.frame(lapply(study, as.numeric))
studyBWT <- study[c(10, 2, 3, 5, 6, 7, 8, 9, 11, 12)]
studyLOW <- study[c(1, 2, 3, 5, 6, 7, 8, 9, 11, 12)]
```

2 Studio del modello di regressione lineare multipla

2.1 Dipendenza tra le variabili

In uno studio di regressione multipla è indispensabile osservare quali sono le variabili che effettivamente incidono sulla variabile obiettivo.
Una volta trovare le variabili di interesse, possiamo costruire il nostro modello di regressione.
Per compiere uno studio approfondito, si è deciso di eseguire tre analisi diverse:

1. un'analisi utilizzando gli algoritmi di scelta del modello già integrati in R;
2. un'analisi utilizzando i grafi non orientati;
3. un'analisi utilizzando i gradi orientati.

2.1.1 Algoritmi di scelta del modello

Per prima cosa, creiamo un modello in cui vengono utilizzate tutte le variabili.

```
attach(studyBWT)
mq0 <- lm(bwt ~ age + lwt + smoke + ptl + ht + ui + ftv + white + black)
summary(mq0)

##
## Call:
## lm(formula = bwt ~ age + lwt + smoke + ptl + ht + ui + ftv +
##     white + black)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.50305 -0.59682  0.07667  0.64928  2.33292
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.12331    0.11981   1.029 0.304772
## age         -0.02594    0.06990  -0.371 0.711012
## lwt          0.18258    0.07278   2.509 0.013007 *
## smoke       -0.48277    0.14602  -3.306 0.001142 **
## ptl         -0.03275    0.06899  -0.475 0.635607
## ht          -0.81297    0.27745  -2.930 0.003830 **
## ui          -0.70772    0.19046  -3.716 0.000271 ***
## ftv         -0.02042    0.06750  -0.303 0.762598
## white        0.48693    0.15737   3.094 0.002290 **
## black       -0.18287    0.21858  -0.837 0.403925
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8918 on 179 degrees of freedom
## Multiple R-squared:  0.2427,Adjusted R-squared:  0.2047
## F-statistic: 6.376 on 9 and 179 DF,  p-value: 7.891e-08
```

Come si può vedere dal summary, questo modello presenta diverse variabili non significative. Queste variabili possono essere quindi eliminate dal modello, in modo da ottenerne uno migliore.

Per svolgere una prima scrematura delle variabili, si è deciso di svolgere degli algoritmi di scelta del modello di direzione *backward*.

```
library(SignifReg)

#Scelta del modello con criterio BIC, direzione backward
mq1 <- SignifReg(mq0, criterion = "BIC", direction = "backward")
summary(mq1)

##
## Call:
## lm(formula = bwt ~ lwt + smoke + ht + ui + white)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54112 -0.63111  0.03589  0.61814  2.22226
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08442    0.10354   0.815 0.415921
## lwt          0.16211    0.06741   2.405 0.017180 *
## smoke       -0.50779    0.13971  -3.635 0.000362 ***
## ht          -0.80145    0.27351  -2.930 0.003819 **
## ui          -0.71658    0.18444  -3.885 0.000143 ***
## white        0.53440    0.13675   3.908 0.000131 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.885 on 183 degrees of freedom
## Multiple R-squared:  0.2377,Adjusted R-squared:  0.2169
## F-statistic: 11.41 on 5 and 183 DF,  p-value: 1.345e-09

#Scelta del modello con criterio AIC, direzione backward
mq2 <- SignifReg(mq0, criterion = "AIC", direction = "backward")
summary(mq2)
```

```
##
## Call:
## lm(formula = bwt ~ lwt + smoke + ht + ui + white)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54112 -0.63111  0.03589  0.61814  2.22226
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08442    0.10354   0.815 0.415921
## lwt          0.16211    0.06741   2.405 0.017180 *
## smoke       -0.50779    0.13971  -3.635 0.000362 ***
## ht          -0.80145    0.27351  -2.930 0.003819 **
## ui          -0.71658    0.18444  -3.885 0.000143 ***
## white        0.53440    0.13675   3.908 0.000131 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.885 on 183 degrees of freedom
## Multiple R-squared:  0.2377,Adjusted R-squared:  0.2169
## F-statistic: 11.41 on 5 and 183 DF,  p-value: 1.345e-09
```

```
#Scelta del modello con criterio p-value, direzione backward
mq3 <- SignifReg(mq0, criterion = "p-value", direction = "backward")
summary(mq3)
```

```
##
## Call:
## lm(formula = bwt ~ lwt + ht + ui + white)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22611 -0.66269  0.06018  0.65437  2.49317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.02259    0.10250  -0.220  0.82583
## lwt          0.17956    0.06944   2.586  0.01049 *
## ht          -0.85514    0.28203  -3.032  0.00278 **
## ui          -0.76467    0.18997  -4.025 8.31e-05 ***
## white        0.37439    0.13370   2.800  0.00565 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9138 on 184 degrees of freedom
```

```
## Multiple R-squared:  0.1827,Adjusted R-squared:  0.1649
## F-statistic: 10.28 on 4 and 184 DF,  p-value: 1.554e-07
```

Come si può osservare, le ricerche che utilizzano i criteri *AIC* e *BIC* hanno portato allo stesso risultato, in cui le uniche variabili rilevanti nel nostro studio sono: **lwt, smoke, ht, ui, white**.
La ricerca con criterio *p-value*, invece, ha riportato un insieme di variabili sottoinsieme di quello precedente.
Per capire quale dei risultati è migliore, possiamo considerare i seguenti aspetti:

1. la variabile eliminata dalla scelta con criterio *p-value* è la variabile **smoke**, la quale ha una significatività elevata nell'altro modello;
2. il p-value della statistica F dell'ultimo modello è maggiore (e quindi meno significativo) rispetto all'altro;
3. rispetto all'ultimo modello, nel primo vi è un aumento significativo dell'R quadro, sia aggiustato (da 0,1649 a 0,2169) che non (da 0,1827 a 0,2377).

Queste considerazioni, ci fanno intuire che il modello più accurato è quello trovato dai criteri *AIC* e *BIC*.

La dipendenza di queste variabili l'una dall'altra ci viene suggerita anche dalla matrice di correlazione parziale.

```
library(gRbase)

S.BWT <- cov.wt(studyBWT, method = "ML")$cov
PC.BWT <- cov2pcor(S.BWT)
round(100*PC.BWT)
```

```
##       bwt age lwt smoke ptl  ht  ui ftv white black
## bwt   100  -3  18   -24  -4 -21 -27  -2    23    -6
## age    -3 100  17   -10  13  -4  -7  18    15    -9
## lwt    18  17 100    -2 -10  27  -2  10    12    29
## smoke -24 -10  -2   100  17  -1  -4  -3    41    14
## ptl    -4  13 -10    17 100   2  19  -3    -5    -1
## ht    -21  -4  27    -1   2 100 -14 -10    -1    -2
## ui    -27  -7  -2    -4  19 -14 100  -3     3    -4
## ftv    -2  18  10    -3  -3 -10  -3 100     5     3
## white  23  15  12    41  -5  -1   3   5   100   -41
## black  -6  -9  29    14  -1  -2  -4   3   -41   100
```

Maggiore è il valore della correlazione in valore assoluto tra due variabili, più queste sono dipendenti.
Come si può osservare dalla matrice la variabile obiettivo **bwt** presenta effettivamente valori elevati di correlazione solo con le variabili **lwt**, **smoke**, **ht**, **ui** e

white. Questo corrobora l'ipotesi di modello ottenuta dalle ricerche svolte. Si vuole inoltre far notare che il valore di correlazione tra **bwt** e **smoke** è uno dei più elevati, altra motivazione a favore della scelta del modello ottenuto tramite i criteri *AIC* e *BIC* rispetto a quello ottenuto con il criterio del *p-value*.

Inoltre, svolgendo gli stessi algoritmi ma in direzione *forward*

```
mq4 <- lm(bwt~1)
summary(mq4)

##
## Call:
## lm(formula = bwt ~ 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.06575 -0.72762  0.04445  0.74383  2.80495
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.076e-17  7.274e-02       0        1
##
## Residual standard error: 1 on 188 degrees of freedom

#Scelta del modello con criterio BIC, direzione forward
mq5 <- SignifReg(mq4, criterion = "BIC", direction = "forward",
                 scope = formula(mq0))
summary(mq5)

##
## Call:
## lm(formula = bwt ~ ui + white + smoke + ht + lwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54112 -0.63111  0.03589  0.61814  2.22226
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08442    0.10354   0.815 0.415921
## ui          -0.71658    0.18444  -3.885 0.000143 ***
## white        0.53440    0.13675   3.908 0.000131 ***
## smoke       -0.50779    0.13971  -3.635 0.000362 ***
## ht          -0.80145    0.27351  -2.930 0.003819 **
## lwt          0.16211    0.06741   2.405 0.017180 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.885 on 183 degrees of freedom
## Multiple R-squared:  0.2377,Adjusted R-squared:  0.2169
## F-statistic: 11.41 on 5 and 183 DF,  p-value: 1.345e-09

#Scelta del modello con criterio AIC, direzione forward
mq6 <- SignifReg(mq4, criterion = "AIC", direction = "forward",
                 scope = formula(mq0))
summary(mq6)

##
## Call:
## lm(formula = bwt ~ ui + white + smoke + ht + lwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54112 -0.63111  0.03589  0.61814  2.22226
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08442    0.10354   0.815 0.415921
## ui          -0.71658    0.18444  -3.885 0.000143 ***
## white        0.53440    0.13675   3.908 0.000131 ***
## smoke       -0.50779    0.13971  -3.635 0.000362 ***
## ht          -0.80145    0.27351  -2.930 0.003819 **
## lwt          0.16211    0.06741   2.405 0.017180 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.885 on 183 degrees of freedom
## Multiple R-squared:  0.2377,Adjusted R-squared:  0.2169
## F-statistic: 11.41 on 5 and 183 DF,  p-value: 1.345e-09

#Scelta del modello con criterio p-value, direzione forward
mq7 <- SignifReg(mq4, criterion = "p-value", direction = "forward",
                 scope = formula(mq0))
summary(mq7)

##
## Call:
## lm(formula = bwt ~ ui + white + smoke + ht + lwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54112 -0.63111  0.03589  0.61814  2.22226
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08442    0.10354   0.815 0.415921
## ui          -0.71658    0.18444  -3.885 0.000143 ***
## white        0.53440    0.13675   3.908 0.000131 ***
## smoke       -0.50779    0.13971  -3.635 0.000362 ***
## ht          -0.80145    0.27351  -2.930 0.003819 **
## lwt          0.16211    0.06741   2.405 0.017180 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.885 on 183 degrees of freedom
## Multiple R-squared:  0.2377,Adjusted R-squared:  0.2169
## F-statistic: 11.41 on 5 and 183 DF,  p-value: 1.345e-09
```

si nota che il modello restituito è sempre lo stesso e corrisponde a quello da noi scelto.

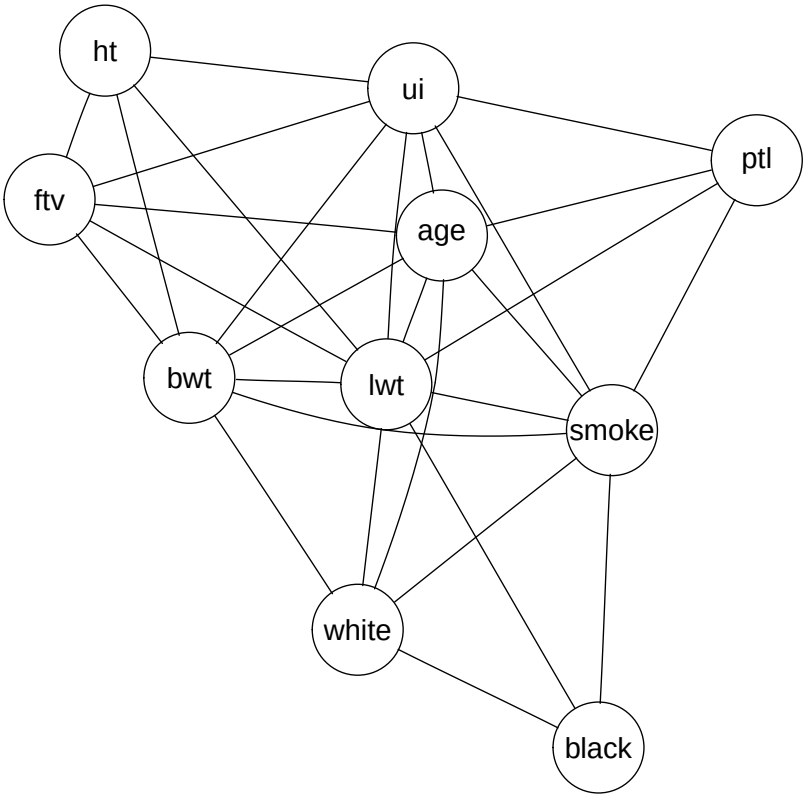
2.1.2 Grafi non orientati

La matrice di correlazione parziale utilizzata in precedenza non è l'unico strumento per osservare la dipendenza tra variabili: un altro metodo è quello di utilizzare un grafo non orientato.

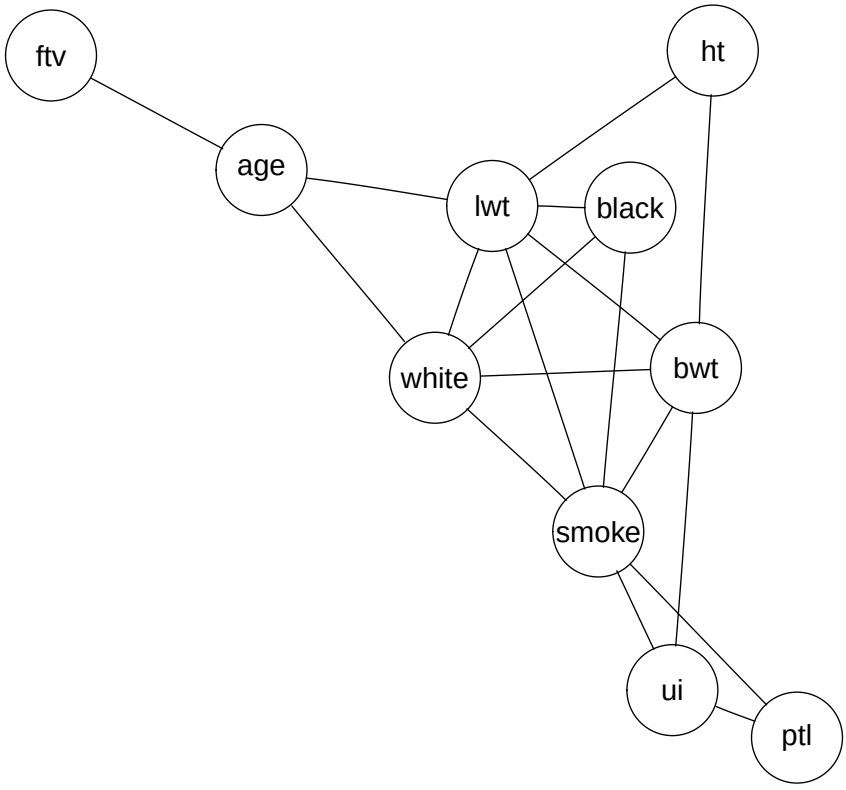
Nei grafi non orientati, i nodi rappresentano le variabili mentre gli archi indicano l'indipendenza condizionata o meno tra esse.
In particolare, date due variabili A e B, queste sono indipendenti condizionatamente data la variabile C se e solo se ogni cammino da A a B passa per C.

Generiamo ora due differenti tipi di grafi, uno basato sul criterio *AIC* e uno sul criterio *BIC* e osserviamo quale dei due si avvicina più alle nostre previsioni.

```
library(gRim)

graph.BWT <- cmod(~.^., data = studyBWT)

#Grafo non orientato, AIC
aic.BWT <- gRbase::stepwise(graph.BWT)
plot(as(aic.BWT, "graphNEL"), "fdp")
```




```
#Grafo non orientato, BIC
bic.BWT <- gRbase::stepwise(graph.BWT, k = log(nrow(studyBWT)))
plot(as(bic.BWT, "graphNEL"), "fdp")
```



Come possiamo osservare, il secondo grafo indica esattamente le stesse dipendenze che abbiamo trovato nella sezione precedente. Nel primo grafo invece vengono anche indicate delle dipendenze tra **bwt** e le variabili **age** e **ftv**.

```
attach(studyBWT)
mq8 <- lm(bwt ~ smoke + ui + ht + lwt + white + age + ftv)
summary(mq8)

##
## Call:
## lm(formula = bwt ~ smoke + ui + ht + lwt + white + age + ftv)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.49604 -0.60602  0.00473  0.63454  2.31916
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08156    0.10430   0.782 0.435221
## smoke       -0.51490    0.14113  -3.648 0.000345 ***
## ui          -0.72197    0.18559  -3.890 0.000141 ***
## ht          -0.81566    0.27648  -2.950 0.003596 **
## lwt          0.16943    0.06913   2.451 0.015197 *
## white        0.54885    0.14066   3.902 0.000134 ***
## age         -0.02502    0.06875  -0.364 0.716378
## ftv         -0.02118    0.06723  -0.315 0.753085
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8891 on 181 degrees of freedom
## Multiple R-squared:  0.2389,Adjusted R-squared:  0.2094
## F-statistic: 8.115 on 7 and 181 DF,  p-value: 1.401e-08
```

Andando a osservare il modello di regressione lineare con tutte le variabili suggerite dal primo grafo si nota un leggero aumento dell'R-quadro, il quale però non è significativo per indicare un miglioramento del modello.
Si può inoltre notare che le due variabili **age** e **ftv** non sono da considerarsi significative e il loro ingresso nel modello comporta anche un aumento del p-value della statistica F (da 1.345e-09 a 1.401e-08) e una diminuzione dell'R-quadro aggiustato (da 0.2169 a 0.2094).

Ciò dimostra che il modello ottenuto in precedenza è ancora da considerarsi migliore e, quindi, possiamo concludere questa sezione notando come i due metodi fino ad ora utilizzati (analisi attraverso la scelta del modello e analisi con i grafi non orientati) restituiscano lo *stesso modello* di regressione lineare.

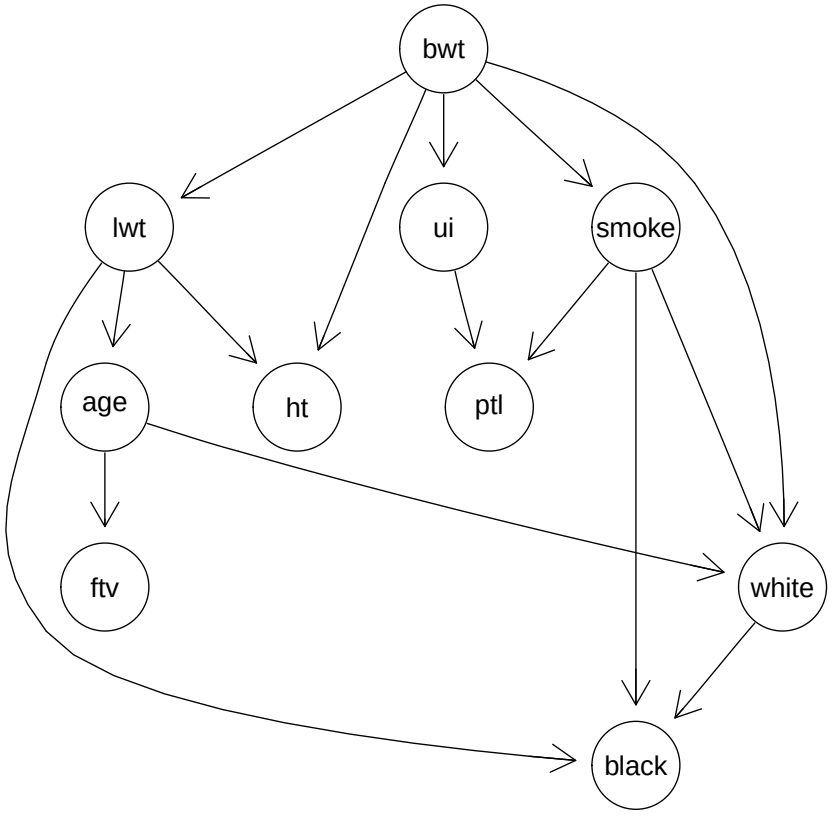
2.1.3 DAG

Infine, svolgiamo l'analisi della dipendenza tra variabili utilizzando i *DAG* (grafi orientati senza cicli).

Così come i grafi non orientati, anche i DAG indicano l'indipendenza condizionata tra le variabili di un dataset.
A differenza del caso precedente, però, gli archi orientati causano anche un ordinamento delle variabili.

Generiamo adesso il DAG.

```
library(bnlearn)

dag1.BWT <- hc(studyBWT)
plot(as(amat(dag1.BWT), "graphNEL"))
```

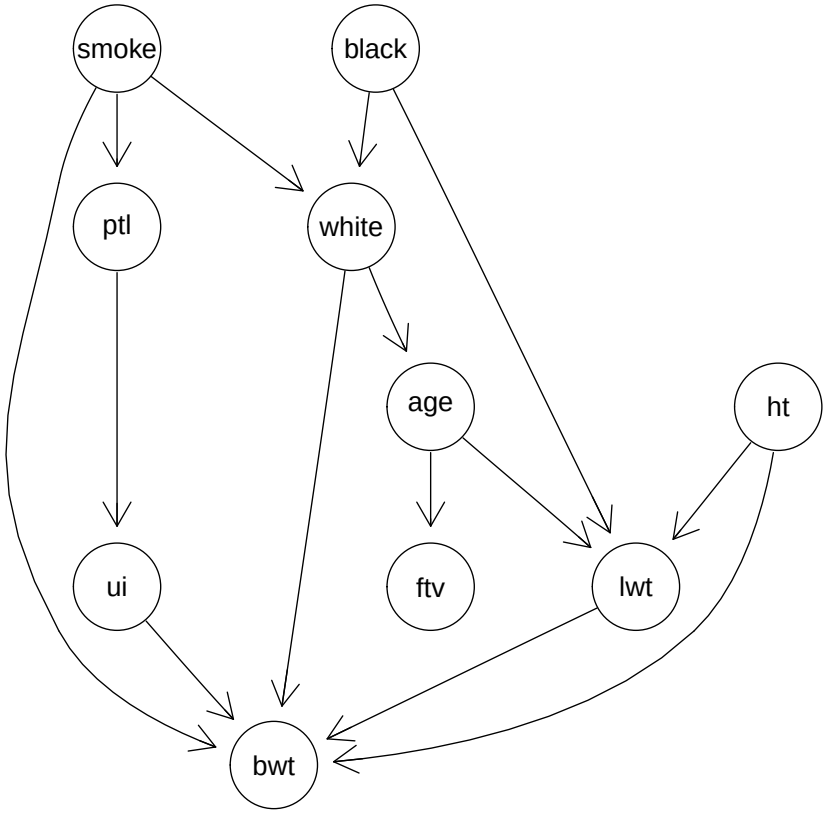


Come si può osservare, il DAG mostra archi uscenti dal nodo **bwt** ma non entranti in esso.
Per evitare che ciò accada, possiamo inserire un ordinamento tra le variabili basato sulla nostra conoscenza pregressa.
In particolare, si è deciso di assegnare tre livelli di ordinamento:

1. age, smoke, white, black, ht;
2. lwt, ptl, ui
3. ftv, bwt

```
library(igraph)

block <- c(3, 1, 2, 1, 2, 1, 2, 3, 1, 1)
blM <- matrix(0, nrow = 10, ncol = 10)
rownames(blM) <- colnames(blM) <- names(studyBWT)
```

```
for (b in 2:3) blM[block==b, block<b] <- 1

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

dag2.BWT <- hc(studyBWT, blacklist = blackL)
plot(as(amat(dag2.BWT), "graphNEL"))
```



Il DAG così ottenuto presenta degli archi coerenti con le nostre conoscenze di background.
Si noti adesso che quello che ci suggerisce questo grafo è che le uniche variabili che influenzano effettivamente il peso del bambino al momento del parto sono: il fatto che la madre fumi (**smoke**), la presenza di irritabilità uterina (**ui**), il peso della madre alla fine dell'ultimo periodo mestruale (**lwt**), la storia familiare di ipertensione (**ht**) e il fatto che la madre sia di razza bianca o meno (**white**); esattamente lo stesso risultato ottenuto nelle altre due analisi.

Quindi, il modello:

```
mq <- mq1
summary(mq)

##
## Call:
## lm(formula = bwt ~ lwt + smoke + ht + ui + white)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54112 -0.63111  0.03589  0.61814  2.22226
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08442    0.10354   0.815 0.415921
## lwt          0.16211    0.06741   2.405 0.017180 *
## smoke       -0.50779    0.13971  -3.635 0.000362 ***
## ht          -0.80145    0.27351  -2.930 0.003819 **
## ui          -0.71658    0.18444  -3.885 0.000143 ***
## white        0.53440    0.13675   3.908 0.000131 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.885 on 183 degrees of freedom
## Multiple R-squared:  0.2377,Adjusted R-squared:  0.2169
## F-statistic: 11.41 on 5 and 183 DF,  p-value: 1.345e-09
```

è quello che riteniamo sia il più accurato per svolgere una regressione lineare multipla su questo dataset.

Osservando il risultato ottenuto, però, è interessante considerare il fatto che la variabile **black** non sia significativa mentre la variabile **white** sì.
Questo può ossere dovuto sia ad un effettiva maggiore significatività della variabile **white**, sia ad un problema intrinseco del dataset.
Nella prossima sezione si compie un'analisi del problema, per essere sicuri di aver considerato il modello giusto.

2.2 Studio della dipendenza del peso del bambino dalla razza della madre

```
attach(studyBWT)
mq9 <- lm(bwt ~ smoke + ui + ht + lwt + black + white)
summary(mq9)
```

```
##
## Call:
## lm(formula = bwt ~ smoke + ui + ht + lwt + black + white)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.52620 -0.59405  0.09201  0.62973  2.23669
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.13047    0.11837   1.102 0.271801
## smoke       -0.48864    0.14186  -3.445 0.000710 ***
## ui          -0.72067    0.18469  -3.902 0.000134 ***
## ht          -0.80250    0.27378  -2.931 0.003810 **
## lwt          0.17787    0.07026   2.532 0.012198 *
## black       -0.17403    0.21611  -0.805 0.421708
## white        0.47743    0.15408   3.099 0.002254 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8858 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
```

Dai risultati ottenuti in precedenza si è notato come la variabile **white** sia (altamente) significativa all'interno del modello di regressione, mentre la variabile **black** non lo sia.
Se proviamo a generare il modello di regressione lineare multipla inserendo anche la variabile **black**, si nota infatti un'abbassamento della significatività di **white**, un'innalzamento del p-value della statistica F e un'abbassamento del R-quadro aggiustato; tutti elementi che portano a escludere l'importanza di questa variabile all'interno dello studio.

Nonostante ciò, non si può escludere la possibilità che questa non significatività della variabile **black** sia data da un problema del dataset stesso.
Infatti, i motivi per cui questa variabile può non essere significativa nel nostro studio, sono due:

1. Il peso delle nascite di figli di madri della razza nera e quella di "altro" sono molto simili tra loro, e quindi è importante distinguere solo tra madri di razza bianca e no; in questo caso l'analisi svolta è corretta.

2. I dati ottenuti presentano una minoranza di madri della razza nera, che causa questo squilibrio e, in questo caso, l'analisi potrebbe non essere corretta.

Per essere sicuri di non trovarci nel secondo caso, si è quindi deciso di svolgere

una breve analisi sulle variabili **white** e **black**, in modo da capire se, effettivamente, la seconda sia significativa o meno.

Come prima cosa, vediamo quanti dati sono presenti per ogni valore della variabile **race**:

```
table(study$race)

##
##  1  2  3
## 96 26 67
```

Come si può vedere le variabili con race = 2 (cioè quelle di **black**) sono decisamente inferiori a quelle di **white** e altro. Per osservare se questo è o meno il motivo per cui vi è uno squilibrio, si può provare a studiare un dataset in cui il numero di madre delle tre razze coincide.
A questo proposito, è opportuno notare che a causa dell'esiguo numero di madri di razza nera, il dataset risultante avrà un numero di righe molto ridotto (solo 78). Per questo motivo, si è deciso di generare 5 diversi dataset e di osservare il modello di regressione lineare multipla risultante per tutti e 5, così da avere un quadro di insieme migliore.

```
studyRace <- list()

for (i in 1:5){
  studyWhite <- studyBWT[sample(which(studyBWT$white==1),26),]
  studyBlack <- studyBWT[sample(which(studyBWT$black==1),26),]
  studyOther <- studyBWT[sample(which(studyBWT$white==0&studyBWT$black==0),26),]
  studyRace[[i]] <- rbind(studyWhite, studyBlack, studyOther)
}

mqrace <- list()
mqrace1 <- list()

for(i in 1:5){
  attach(studyRace[[i]])
  mqrace[[i]] <- lm(bwt ~ smoke + ui + ht + lwt + black + white)
  mqrace1[[i]] <- lm(bwt ~ smoke + ui + ht + lwt + white)
}

summary(mqrace[[1]])

##
## Call:
## lm(formula = bwt ~ smoke + ui + ht + lwt + black + white)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67260 -0.65759  0.05136  0.64399  2.34200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.2294     0.2003   1.145  0.25592
## smoke        -0.5408     0.2110  -2.563  0.01250 *
## ui           -0.9883     0.2939  -3.363  0.00125 **
## ht           -0.9177     0.4323  -2.123  0.03726 *
## lwt           0.2913     0.1018   2.861  0.00554 **
## black        -0.2718     0.2676  -1.015  0.31334
## white         0.2984     0.2651   1.126  0.26398
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8816 on 71 degrees of freedom
## Multiple R-squared:  0.3226,Adjusted R-squared:  0.2653
## F-statistic: 5.635 on 6 and 71 DF,  p-value: 7.901e-05
```

```
summary(mqrace[[2]])
```

```
##
## Call:
## lm(formula = bwt ~ smoke + ui + ht + lwt + black + white)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9041 -0.5879  0.1014  0.6249  2.0301
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.09124    0.19317  -0.472  0.63816
## smoke       -0.37629    0.23070  -1.631  0.10730
## ui          -0.75383    0.27782  -2.713  0.00835 **
## ht          -0.70303    0.40444  -1.738  0.08649 .
## lwt          0.12624    0.11655   1.083  0.28241
## black        0.02551    0.27902   0.091  0.92742
## white        0.89426    0.28078   3.185  0.00215 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8987 on 71 degrees of freedom
## Multiple R-squared:  0.2342,Adjusted R-squared:  0.1695
## F-statistic: 3.619 on 6 and 71 DF,  p-value: 0.003447
```

```
summary(mqrace[[3]])
```

```
##
## Call:
## lm(formula = bwt ~ smoke + ui + ht + lwt + black + white)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6719 -0.4958  0.1137  0.5390  1.4315
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.23682    0.18131   1.306   0.1957
## smoke       -0.34992    0.21241  -1.647   0.1039
## ui          -0.52771    0.27036  -1.952   0.0549 .
## ht          -0.78666    0.54778  -1.436   0.1554
## lwt          0.19768    0.10513   1.880   0.0642 .
## black       -0.36883    0.26270  -1.404   0.1647
## white       -0.04302    0.25967  -0.166   0.8689
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8475 on 71 degrees of freedom
## Multiple R-squared:  0.1703,Adjusted R-squared:  0.1002
## F-statistic: 2.428 on 6 and 71 DF,  p-value: 0.03422
```

```
summary(mqrace[[4]])
```

```
##
## Call:
## lm(formula = bwt ~ smoke + ui + ht + lwt + black + white)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60050 -0.56971  0.04761  0.52983  1.80524
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.2317     0.1817   1.275 0.206441
## smoke        -0.3641     0.2211  -1.646 0.104104
## ui           -0.6610     0.2568  -2.574 0.012127 *
## ht           -1.4161     0.3817  -3.710 0.000408 ***
## lwt           0.2865     0.1100   2.603 0.011237 *
## black        -0.3196     0.2609  -1.225 0.224621
## white         0.3530     0.2600   1.358 0.178804
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.8319 on 71 degrees of freedom
## Multiple R-squared:  0.2727,Adjusted R-squared:  0.2112
## F-statistic: 4.436 on 6 and 71 DF,  p-value: 0.0007262
```

```
summary(mqrace[[5]])
```

```
##
## Call:
## lm(formula = bwt ~ smoke + ui + ht + lwt + black + white)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.17685 -0.47310  0.03815  0.62415  1.46819
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.25133    0.18605   1.351   0.1810
## smoke       -0.54706    0.21501  -2.544   0.0131 *
## ui          -0.46456    0.28668  -1.620   0.1096
## ht          -0.33105    0.43650  -0.758   0.4507
## lwt          0.09874    0.10571   0.934   0.3535
## black       -0.31240    0.26453  -1.181   0.2416
## white        0.35405    0.26070   1.358   0.1787
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8696 on 71 degrees of freedom
## Multiple R-squared:  0.1888,Adjusted R-squared:  0.1202
## F-statistic: 2.754 on 6 and 71 DF,  p-value: 0.01831
```

Come possiamo osservare, la variabile **black** continua a non essere particolarmente significativa (se non in rari casi).
Possiamo quindi concludere la nostra analisi sostenendo che la variabile obiettivo **bwt** è indipendente da **black** condizionatamente alle altre (e, in particolare, a **white** secondo il vecchio DAG) e quindi il modello da considerare corretto è quello trovato nella sezione precedente.

Infine, un ultimo strumento che possiamo utilizzare per osservare la maggior accuratezza del modello senza la variabile **black** è il *test del rapporto di verosimiglianza*.

```
library(lmtest)

lrtest(mqrace1[[1]], mqrace[[1]])

## Likelihood ratio test
```

```
##
## Model 1: bwt ~ smoke + ui + ht + lwt + white
## Model 2: bwt ~ smoke + ui + ht + lwt + black + white
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -97.741
## 2   8 -97.179  1 1.1246     0.2889

lrtest(mqrace1[[2]], mqrace[[2]])

## Likelihood ratio test
##
## Model 1: bwt ~ smoke + ui + ht + lwt + white
## Model 2: bwt ~ smoke + ui + ht + lwt + black + white
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -98.688
## 2   8 -98.683  1 0.0092     0.9237

lrtest(mqrace1[[3]], mqrace[[3]])

## Likelihood ratio test
##
## Model 1: bwt ~ smoke + ui + ht + lwt + white
## Model 2: bwt ~ smoke + ui + ht + lwt + black + white
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -95.171
## 2   8 -94.103  1 2.1362     0.1439

lrtest(mqrace1[[4]], mqrace[[4]])

## Likelihood ratio test
##
## Model 1: bwt ~ smoke + ui + ht + lwt + white
## Model 2: bwt ~ smoke + ui + ht + lwt + black + white
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -93.470
## 2   8 -92.654  1 1.6314     0.2015

lrtest(mqrace1[[5]], mqrace[[5]])

## Likelihood ratio test
##
## Model 1: bwt ~ smoke + ui + ht + lwt + white
## Model 2: bwt ~ smoke + ui + ht + lwt + black + white
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -96.874
## 2   8 -96.116  1 1.5173      0.218
```

Svolgendo un test per ogni dataset creato nella scorsa iterazione, possiamo osservare come non vi siano valori di p-value particolarmente significativi: questo significa che in nessun caso rifiutiamo l'ipotesi nulla e, quindi, non ci sono motivazioni per scegliere un modello più complesso al posto di quello dove la variabile **black** non è presente.

2.3 Svolgimento di alcuni test per confermare o meno l'ipotesi del modello

Dallo studio fino ad ora svolto, abbiamo l'ipotesi che il modello migliore sia:

```
summary(mq)

##
## Call:
## lm(formula = bwt ~ lwt + smoke + ht + ui + white)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54112 -0.63111  0.03589  0.61814  2.22226
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08442    0.10354   0.815 0.415921
## lwt          0.16211    0.06741   2.405 0.017180 *
## smoke       -0.50779    0.13971  -3.635 0.000362 ***
## ht          -0.80145    0.27351  -2.930 0.003819 **
## ui          -0.71658    0.18444  -3.885 0.000143 ***
## white        0.53440    0.13675   3.908 0.000131 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.885 on 183 degrees of freedom
## Multiple R-squared:  0.2377,Adjusted R-squared:  0.2169
## F-statistic: 11.41 on 5 and 183 DF,  p-value: 1.345e-09
```

Ciò che vogliamo fare in questa sezione è uno studio finale su questo modello, per osservare quanto questo risulta corretto alla luce di diversi test.

2.3.1 Analisi dei residui

In un problema di regressione lineare multipla si ha come ipotesi che gli errori del modello siano distribuiti come una distribuzione normale standard.
Un modo per osservare se questa ipotesi è rispettata o meno, è quello di osservare i residui del modello trovato.

```
e <- mq$residuals

#somma dei residui
sum(e)

## [1] 1.052977e-14

#mediata dei residui
median(e)

## [1] 0.03589176
```

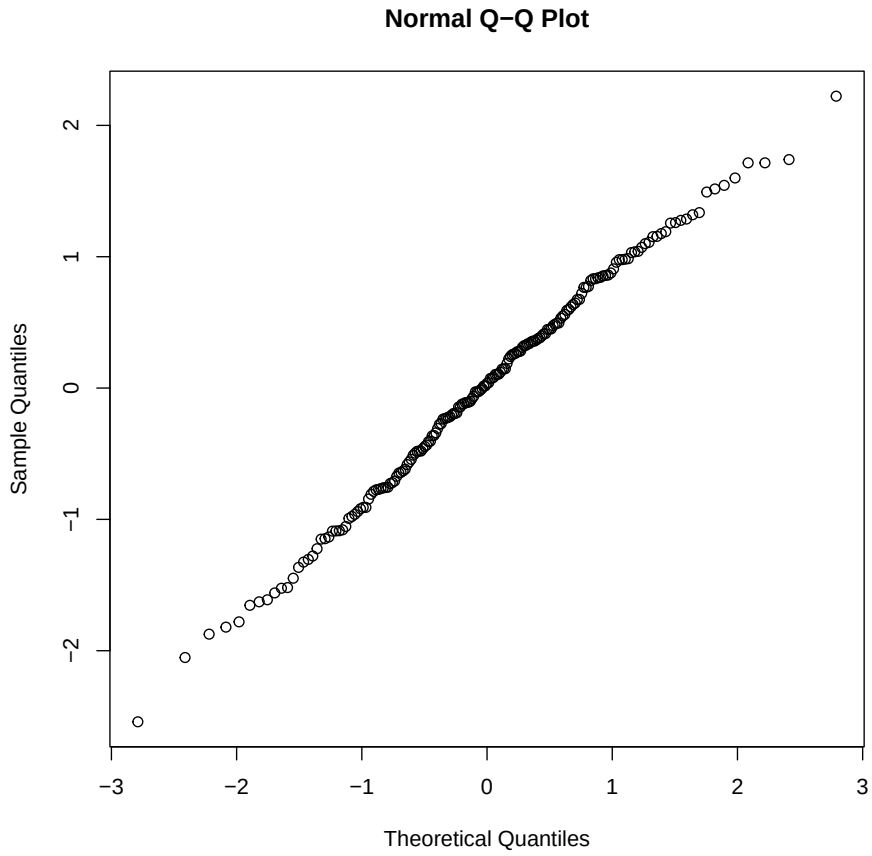
Come si può osservare, la somma dei residui (molto vicina a zero) ci fa pensare che questi abbiano una distribuzione normale standard.
Inoltre, anche il valore della mediana è significativo: in una distribuzione normale standard la mediana coincide con la media, la quale è uguale a zero. Il valore della mediana di questi residui non è così vicino a zero come la loro somma, ma considerando il range di valori di questo vettore:

```
max(e)

## [1] 2.222258

min(e)

## [1] -2.541117
```

si nota che si raggiungono anche valori molto "lontani" da zero. Di conseguenza l'ordine di 10^{-2} della mediana ci fa intuire una buona approssimazione alla distribuzione normale standard.

Andiamo adesso a osservare anche il QQ plot tra i residui e i quantili di una distribuzione normale tenendo a mente che più i punti approssimano la retta bisettrice del primo quadrante del grafico, più allora l'approssimazione dei residui a una distribuzione normale standard è corretta.

```
qqnorm(e)
```



Il grafico conferma le nostre ipotesi.

2.3.2 Intervalli di confidenza

Un'ulteriore analisi che possiamo compiere è quella sugli intervalli di confidenza dei parametri trovati:

```
#Intervalli di confidenza al 95%
confint(mq)

##                   2.5 %     97.5 %
## (Intercept) -0.11985829  0.2886990
## lwt          0.02910559  0.2951198
## smoke       -0.78344804 -0.2321365
## ht          -1.34109292 -0.2618009
## ui          -1.08047832 -0.3526807
## white        0.26459203  0.8042157
```

Osservare gli intervalli di confidenza può essere utile per confermare ulteriormente la significatività delle variabili scelte nel modello.
Come si può notare, infatti, nessuna delle stime dei coefficienti delle variabili ha al proprio interno il valore 0: questo significa che la variabile è significativa per lo studio della variabile obiettivo, così come avevamo già osservato in precedenza.

Se andiamo a osservare gli intervalli di confidenza di un modello che abbiamo scartato, invece, è possibile trovare intervalli che contengano il valore 0.

```
confint(mq0)
```

```
##                    2.5 %     97.5 %
## (Intercept) -0.11311655  0.3597428
## age         -0.16388078  0.1120005
## lwt          0.03896455  0.3262039
## smoke       -0.77090512 -0.1946397
## ptl         -0.16887979  0.1033881
## ht          -1.36046275 -0.2654720
## ui          -1.08355555 -0.3318884
## ftv         -0.15362207  0.1127795
## white        0.17640048  0.7974617
## black       -0.61419747  0.2484602
```

2.4 Regressione lineare multipla con polinomi non di primo grado

Il modello ottenuto dallo studio fino ad ora svolto è:

```
attach(studyBWT)
poly1 <- lm(bwt ~ lwt + smoke + ht + ui + white)
summary(poly1)
```

```
##
## Call:
## lm(formula = bwt ~ lwt + smoke + ht + ui + white)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54112 -0.63111  0.03589  0.61814  2.22226
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08442    0.10354   0.815 0.415921
## lwt          0.16211    0.06741   2.405 0.017180 *
## smoke       -0.50779    0.13971  -3.635 0.000362 ***
```

```
## ht          -0.80145    0.27351  -2.930 0.003819 **
## ui          -0.71658    0.18444  -3.885 0.000143 ***
## white        0.53440    0.13675   3.908 0.000131 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.885 on 183 degrees of freedom
## Multiple R-squared:  0.2377,Adjusted R-squared:  0.2169
## F-statistic: 11.41 on 5 and 183 DF,  p-value: 1.345e-09
```

Il modello così ottenuto è un modello di regressione lineare multipla con un polinomio di *primo grado*.
Per completare lo studio, si è pensato di andare a osservare cosa accade nel caso in cui si provi a utilizzare polinomi di grado superiore al primo.
Visto che l'unica variabile non binaria presente nel modello è la variabile **lwt**, sarà solo questa a influire sul grado del polinomio.

```
attach(studyBWT)

#Polinomio di secondo grado
poly2 <- update(poly1, .~. + I(lwt^2))
summary(poly2)

##
## Call:
## lm(formula = bwt ~ lwt + smoke + ht + ui + white + I(lwt^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.52635 -0.63059  0.04985  0.60621  2.22202
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.06908    0.11118   0.621 0.535177
## lwt          0.13836    0.09155   1.511 0.132453
## smoke       -0.51839    0.14272  -3.632 0.000365 ***
## ht          -0.81466    0.27630  -2.948 0.003612 **
## ui          -0.72531    0.18626  -3.894 0.000138 ***
## white        0.54385    0.13926   3.905 0.000132 ***
## I(lwt^2)     0.01691    0.04398   0.385 0.701028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.887 on 182 degrees of freedom
## Multiple R-squared:  0.2383,Adjusted R-squared:  0.2132
## F-statistic:  9.49 on 6 and 182 DF,  p-value: 4.546e-09
```

```r
#Polinomio di terzo grado
poly3 <- update(poly2, .~. + I(lwt^3))
summary(poly3)
```

```
##
## Call:
## lm(formula = bwt ~ lwt + smoke + ht + ui + white + I(lwt^2) +
##     I(lwt^3))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64708 -0.56568  0.02487  0.58262  2.14173
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.13118    0.11502   1.140  0.25561
## lwt          0.06887    0.09784   0.704  0.48238
## smoke       -0.51034    0.14175  -3.600  0.00041 ***
## ht          -0.81738    0.27429  -2.980  0.00328 **
## ui          -0.67713    0.18660  -3.629  0.00037 ***
## white        0.55473    0.13836   4.009 8.89e-05 ***
## I(lwt^2)    -0.13580    0.09079  -1.496  0.13645
## I(lwt^3)     0.05375    0.02802   1.918  0.05663 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8806 on 181 degrees of freedom
## Multiple R-squared:  0.2535,Adjusted R-squared:  0.2246
## F-statistic:  8.78 on 7 and 181 DF,  p-value: 2.791e-09
```

```r
#Polinomio di quarto grado
poly4 <- update(poly3, .~. + I(lwt^4))
summary(poly4)
```

```
##
## Call:
## lm(formula = bwt ~ lwt + smoke + ht + ui + white + I(lwt^2) +
##     I(lwt^3) + I(lwt^4))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.65727 -0.57234  0.01714  0.60836  2.14712
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.142902   0.116345   1.228 0.220954
```

```
## lwt          0.161300   0.162432   0.993 0.322028
## smoke       -0.516236   0.142179  -3.631 0.000368 ***
## ht          -0.818717   0.274669  -2.981 0.003273 **
## ui          -0.682402   0.186998  -3.649 0.000344 ***
## white        0.557905   0.138617   4.025 8.39e-05 ***
## I(lwt^2)    -0.143677   0.091583  -1.569 0.118447
## I(lwt^3)    -0.006545   0.089059  -0.073 0.941493
## I(lwt^4)     0.015394   0.021579   0.713 0.476520
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8818 on 180 degrees of freedom
## Multiple R-squared:  0.2556,Adjusted R-squared:  0.2225
## F-statistic: 7.725 on 8 and 180 DF,  p-value: 6.937e-09
```

```
#Polinomio di quinto grado
poly5 <- update(poly4, .~. + I(lwt^5))
summary(poly5)
```

```
##
## Call:
## lm(formula = bwt ~ lwt + smoke + ht + ui + white + I(lwt^2) +
##     I(lwt^3) + I(lwt^4) + I(lwt^5))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.67969 -0.56872  0.00404  0.61345  2.12684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.18224    0.12687   1.436 0.152641
## lwt          0.20324    0.17123   1.187 0.236817
## smoke       -0.51600    0.14233  -3.625 0.000376 ***
## ht          -0.83431    0.27569  -3.026 0.002841 **
## ui          -0.69574    0.18798  -3.701 0.000285 ***
## white        0.55409    0.13885   3.990 9.61e-05 ***
## I(lwt^2)    -0.27050    0.18633  -1.452 0.148327
## I(lwt^3)    -0.03031    0.09419  -0.322 0.747986
## I(lwt^4)     0.07180    0.07531   0.953 0.341673
## I(lwt^5)    -0.01125    0.01439  -0.782 0.435344
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8827 on 179 degrees of freedom
## Multiple R-squared:  0.2581,Adjusted R-squared:  0.2208
## F-statistic:  6.92 on 9 and 179 DF,  p-value: 1.544e-08
```

Osservando i risultati ottenuti, si nota che aumentando il grado del polinomio il modello di regressione lineare multipla ottenuto peggiora nettamente rispetto al precedente, sia in termini di significatività della variabile **lwt** e delle sue potenze, sia per quanto riguarda la statistica F, sia per il valore dell'R-quadro aggiustato.
L'unica eccezione è il polinomio di terzo grado. Esso infatti presenta alcune caratteristiche particolari:

1. il suo R-quadro aggiustato e il p-value della statistica F sono migliori rispetto al polinomio di grado precedente;

2. la sua statistica F è meno significativa di quella del modello con il polinomio di grado 1, ma l'R-quadro aggiustato è maggiore, indicando un possibile miglioramento

3. la variabile **lwt** alla terza è leggermente significativa, a differenza degli altri polinomi in cui qualsiasi potenza di **lwt** perde completamente la sua significatività.

Per questo motivo, si fa uno studio più approfondito sulla possibilità che il modello da trovare contenga effettivamente la variabile **lwt** alla terza al suo interno.
Per fare ciò, si studiano tutti i possibili polinomi dove questo accade.

```
attach(studyBWT)

#Polinomio con variabile lwt solo alla terza
poly3v1 <- lm(bwt ~ I(lwt^3) + smoke + ht + ui + white)
summary(poly3v1)

##
## Call:
## lm(formula = bwt ~ I(lwt^3) + smoke + ht + ui + white)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.52889 -0.60193  0.03677  0.58511  2.17239
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.04800    0.10369   0.463 0.644011
## I(lwt^3)     0.02425    0.00926   2.619 0.009561 **
## smoke       -0.55234    0.13919  -3.968 0.000104 ***
## ht          -0.82464    0.27366  -3.013 0.002950 **
## ui          -0.74182    0.18286  -4.057 7.36e-05 ***
## white        0.58483    0.13581   4.306 2.70e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8824 on 183 degrees of freedom
## Multiple R-squared:  0.242,Adjusted R-squared:  0.2213
## F-statistic: 11.69 on 5 and 183 DF,  p-value: 8.209e-10

#Polinomio con variabile lwt alla terza e alla seconda
poly3v2 <- update(poly3v1, .~. + I(lwt^2))
summary(poly3v2)

##
## Call:
## lm(formula = bwt ~ I(lwt^3) + smoke + ht + ui + white + I(lwt^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64044 -0.59049  0.04073  0.62766  2.11849
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.12029    0.11382   1.057 0.291980
## I(lwt^3)     0.06106    0.02599   2.349 0.019897 *
## smoke       -0.52609    0.13978  -3.764 0.000226 ***
## ht          -0.80244    0.27309  -2.938 0.003726 **
## ui          -0.69141    0.18524  -3.733 0.000253 ***
## white        0.57367    0.13553   4.233 3.66e-05 ***
## I(lwt^2)    -0.13729    0.09064  -1.515 0.131606
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8793 on 182 degrees of freedom
## Multiple R-squared:  0.2514,Adjusted R-squared:  0.2268
## F-statistic: 10.19 on 6 and 182 DF,  p-value: 1.034e-09

#Polinomio con variabile lwt alla terza e alla prima
poly3v3 <- update(poly3v1, .~. + lwt)
summary(poly3v3)

##
## Call:
## lm(formula = bwt ~ I(lwt^3) + smoke + ht + ui + white + lwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.53713 -0.62353  0.03714  0.59207  2.19617
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.06024    0.10514   0.573 0.567423
## I(lwt^3)     0.01701    0.01352   1.258 0.210037
## smoke       -0.53551    0.14122  -3.792 0.000203 ***
## ht          -0.84006    0.27480  -3.057 0.002572 **
## ui          -0.72628    0.18431  -3.941 0.000116 ***
## white        0.56484    0.13866   4.074  6.9e-05 ***
## lwt          0.07226    0.09814   0.736 0.462478
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8835 on 182 degrees of freedom
## Multiple R-squared:  0.2443,Adjusted R-squared:  0.2193
## F-statistic: 9.804 on 6 and 182 DF,  p-value: 2.334e-09
```

Osservando i risultati si osserva che il polinomio in cui la variabile lwt compare solo alla terza è nettamente migliore del nostro modello iniziale.
Esso infatti denota:

1. un aumento della significatività della variabile **lwt** (da * a **);
2. una diminuizione del p-value della statistica F (da 1.345e-09 a 8.209e-10);
3. un aumento dell'R-quadro aggiustato (da 0.2169 a 0.2213).

3 Breve studio del modello di regressione logistica multipla

Essendo la variabile obiettivo di questo studio (**low**) generata da una dicotomizzazione della variabile obiettivo dello studio precedente (**bwt**), ci si aspetta che il modello di regressione di questo studio dipenda dalle stesse variabili trovate in precedenza o da un loro sottoinsieme.

```
attach(studyLOW)

fit <- glm(low ~ lwt + smoke + ht + ui + white, family = binomial)
summary(fit)

##
## Call:
## glm(formula = low ~ lwt + smoke + ht + ui + white, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7582  -0.8357  -0.5349   1.0000   2.1762
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.0796     0.2688  -4.016 5.91e-05 ***
## lwt          -0.4675     0.1974  -2.368  0.01788 *
## smoke         1.0912     0.3866   2.822  0.00477 **
## ht            1.8533     0.6882   2.693  0.00709 **
## ui            0.8926     0.4495   1.986  0.04708 *
## white        -1.0648     0.3881  -2.743  0.00608 **
## ---
## 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: 204.77  on 183  degrees of freedom
## AIC: 216.77
##
## Number of Fisher Scoring iterations: 4
```

Il modello ottenuto utilizzando le stesse variabili trovate durante lo studio precedente pare essere accurato.
Inoltre, osservando i coefficienti stimati in questo studio e in quello precedente:

```
BWT <- poly1$coefficients
LOW <- fit$coefficients
cf <- data.frame(BWT, LOW)
cf

##                      BWT        LOW
## (Intercept)  0.08442035 -1.0795840
## lwt          0.16211269 -0.4674625
## smoke       -0.50779225  1.0911548
## ht          -0.80144692  1.8532869
## ui          -0.71657950  0.8925557
## white        0.53440385 -1.0647689
```

Si può notare che da un punto di vista di segno c'è coerenza: le variabili che aumentano il valore di **bwt** diminuiscono **low** e viceversa. Questo accade perché **low** ha un valore "opposto" a **bwt**.

3.1 Scelta delle variabili e confronto dei modelli utilizzando le tabelle di confusione

Eseguiamo adesso gli stessi passaggi eseguiti nello scorso studio, per osservare se l'insieme di variabili in questo modello può essere ridotto.
Come metodo di confronto, si è deciso di utilizzare le *tabelle di confusione*: tabelle cioè in cui si evidenzia il numero di veri positivi (TP), falsi positivi (FP), falsi negativi (FN) e veri negativi (TN) predetti dal modello.
Per ogni tabella, si è deciso poi di calcolare la precisione e il recupero, dati dalle formule matematiche:
$precision = \frac{TP}{TP+FP}$
$recall = \frac{TP}{TP+FN}$
Considerando la natura del problema (un problema medico, in cui è più importante diminuire i falsi negativi più che i falsi positivi), maggiore sarà il recupero, più consideremo il modello accurato.
Per calcolare e mostrare a schermo la *matrice di confusione*, il *recupero* e la *precisione* di un modello, si è scritta la seguente funzione:

```
fit.matrix <- function(model){
  result <- table(fit$mode$low, fitted(model) > 0.5)
  rownames(result) <- c("actual 0", "actual 1")
  colnames(result) <- c("predicted 0", "predicted 1")

  return(result)
}

fit.recall.precision <- function(matrix){
```

```
  rec = matrix[2,2]/(matrix[2,2]+matrix[2,1])
  prec = matrix[2,2]/(matrix[2,2]+matrix[1,2])
  recprec <- data.frame(rec, prec)
  colnames(recprec) <- c("recall", "precision")

  return(recprec)
}
```

Per quanto riguarda il modello già trovato, si ha:

```
m <- fit.matrix(fit)
rp <- fit.recall.precision(m)

m

##
##            predicted 0 predicted 1
##   actual 0         117          13
##   actual 1          38          21

rp

##      recall precision
## 1 0.3559322 0.6176471
```

Di conseguenza, il *recupero* non è molto elevato: infatti, solo nel 35% dei casi i bambini che nascono con un peso basso vengono effettivamente diagnosticati. Vediamo se riusciamo a trovare un modello con un *recupero* migliore.

3.1.1 Algoritmi di scelta del modello

Per semplicità, in questa sezione utilizziamo solo due algoritmi con criterio *backward*.

```
library(Rcmdr)
attach(studyLOW)

fit0 <- glm(low ~ age + lwt + smoke + ptl + ht + ui + ftv + white + black,
            family = binomial)

#Scelta del modello con criterio BIC, direzione backward
fit1 <- step(fit0, k = log(length(low)), direction = "backward")

## Start:  AIC=253.7
## low ~ age + lwt + smoke + ptl + ht + ui + ftv + white + black
##
```

```
##         Df Deviance    AIC
## - ftv    1   201.43 248.60
## - black  1   201.81 248.99
## - age    1   201.93 249.11
## - ptl    1   203.83 251.01
## - ui     1   204.03 251.20
## - white  1   205.37 252.55
## <none>       201.28 253.70
## - lwt    1   206.80 253.98
## - smoke  1   206.91 254.08
## - ht     1   208.81 255.98
##
## Step:  AIC=248.6
## low ~ age + lwt + smoke + ptl + ht + ui + white + black
##
##         Df Deviance    AIC
## - black  1   201.98 243.91
## - age    1   201.99 243.92
## - ptl    1   203.95 245.88
## - ui     1   204.11 246.04
## - white  1   205.38 247.32
## <none>       201.43 248.60
## - lwt    1   206.81 248.74
## - smoke  1   206.92 248.85
## - ht     1   208.81 250.75
##
## Step:  AIC=243.91
## low ~ age + lwt + smoke + ptl + ht + ui + white
##
##         Df Deviance    AIC
## - age    1   202.62 239.31
## - ptl    1   204.44 241.14
## - ui     1   204.52 241.21
## - lwt    1   206.81 243.50
## <none>       201.98 243.91
## - smoke  1   208.37 245.06
## - white  1   208.77 245.46
## - ht     1   209.24 245.93
##
## Step:  AIC=239.31
## low ~ lwt + smoke + ptl + ht + ui + white
##
##         Df Deviance    AIC
## - ptl    1   204.77 236.22
## - ui     1   205.39 236.84
```

```
## <none>       202.62 239.31
## - lwt    1   208.12 239.57
## - smoke  1   209.35 240.80
## - ht     1   209.99 241.44
## - white  1   210.31 241.76
##
## Step:  AIC=236.22
## low ~ lwt + smoke + ht + ui + white
##
##         Df Deviance    AIC
## - ui     1   208.66 234.87
## <none>       204.77 236.22
## - lwt    1   211.17 237.38
## - ht     1   212.37 238.58
## - white  1   212.83 239.03
## - smoke  1   213.15 239.36
##
## Step:  AIC=234.87
## low ~ lwt + smoke + ht + white
##
##         Df Deviance    AIC
## <none>       208.66 234.87
## - ht     1   215.38 236.35
## - lwt    1   216.38 237.35
## - white  1   216.86 237.82
## - smoke  1   217.74 238.71
```

```
summary(fit1)
```

```
##
## Call:
## glm(formula = low ~ lwt + smoke + ht + white, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7927  -0.8766  -0.5570   1.0358   2.1202
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.9392     0.2564  -3.664 0.000249 ***
## lwt          -0.5113     0.1984  -2.576 0.009982 **
## smoke         1.1176     0.3816   2.929 0.003405 **
## ht            1.7402     0.6894   2.524 0.011591 *
## white        -1.0599     0.3828  -2.769 0.005627 **
## ---
## 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: 208.66  on 184  degrees of freedom
## AIC: 218.66
##
## Number of Fisher Scoring iterations: 4
```

```
#Scelta del modello con criterio AIC, direzione backward
fit2 <- step(fit0, criterion = "AIC", direction = "backward")
```

```
## Start:  AIC=221.28
## low ~ age + lwt + smoke + ptl + ht + ui + ftv + white + black
##
##         Df Deviance    AIC
## - ftv    1   201.43 219.43
## - black  1   201.81 219.81
## - age    1   201.93 219.93
## <none>       201.28 221.28
## - ptl    1   203.83 221.83
## - ui     1   204.03 222.03
## - white  1   205.37 223.37
## - lwt    1   206.80 224.80
## - smoke  1   206.91 224.91
## - ht     1   208.81 226.81
##
## Step:  AIC=219.43
## low ~ age + lwt + smoke + ptl + ht + ui + white + black
##
##         Df Deviance    AIC
## - black  1   201.98 217.98
## - age    1   201.99 217.99
## <none>       201.43 219.43
## - ptl    1   203.95 219.95
## - ui     1   204.11 220.11
## - white  1   205.38 221.38
## - lwt    1   206.81 222.81
## - smoke  1   206.92 222.92
## - ht     1   208.81 224.81
##
## Step:  AIC=217.98
## low ~ age + lwt + smoke + ptl + ht + ui + white
##
##         Df Deviance    AIC
## - age    1   202.62 216.62
```

```
## <none>       201.98 217.98
## - ptl    1   204.44 218.44
## - ui     1   204.52 218.52
## - lwt    1   206.81 220.81
## - smoke  1   208.37 222.37
## - white  1   208.77 222.77
## - ht     1   209.24 223.24
##
## Step:  AIC=216.62
## low ~ lwt + smoke + ptl + ht + ui + white
##
##         Df Deviance    AIC
## <none>       202.62 216.62
## - ptl    1   204.77 216.77
## - ui     1   205.39 217.39
## - lwt    1   208.12 220.12
## - smoke  1   209.35 221.35
## - ht     1   209.99 221.99
## - white  1   210.31 222.31
```

```
summary(fit2)
```

```
##
## Call:
## glm(formula = low ~ lwt + smoke + ptl + ht + ui + white, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8762  -0.8176  -0.5218   0.9622   2.1880
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.0358     0.2716  -3.814 0.000137 ***
## lwt          -0.4383     0.1987  -2.205 0.027432 *
## smoke         0.9992     0.3927   2.544 0.010949 *
## ptl           0.2438     0.1685   1.447 0.147872
## ht            1.8370     0.6929   2.651 0.008025 **
## ui            0.7712     0.4591   1.680 0.092977 .
## white        -1.0482     0.3904  -2.685 0.007258 **
## ---
## 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: 202.62  on 182  degrees of freedom
```

```
## AIC: 216.62
##
## Number of Fisher Scoring iterations: 4
```

Si sono ottenuti due diversi modelli:

1. il criterio *BIC* ha restituito un modello in cui viene eliminata la variabile **ui** rispetto al modello da noi assunto corretto;
2. il criterio *AIC* ha generato un modello in cui si è aggiunta la variabile **ptl** rispetto al modello da noi assunto corretto; si noti che questa variabile non ha significatività.

Studiamo quindi la *matrice di confusione*, il *recupero* e la *precisione* dei due nuovi modelli ottenuti:

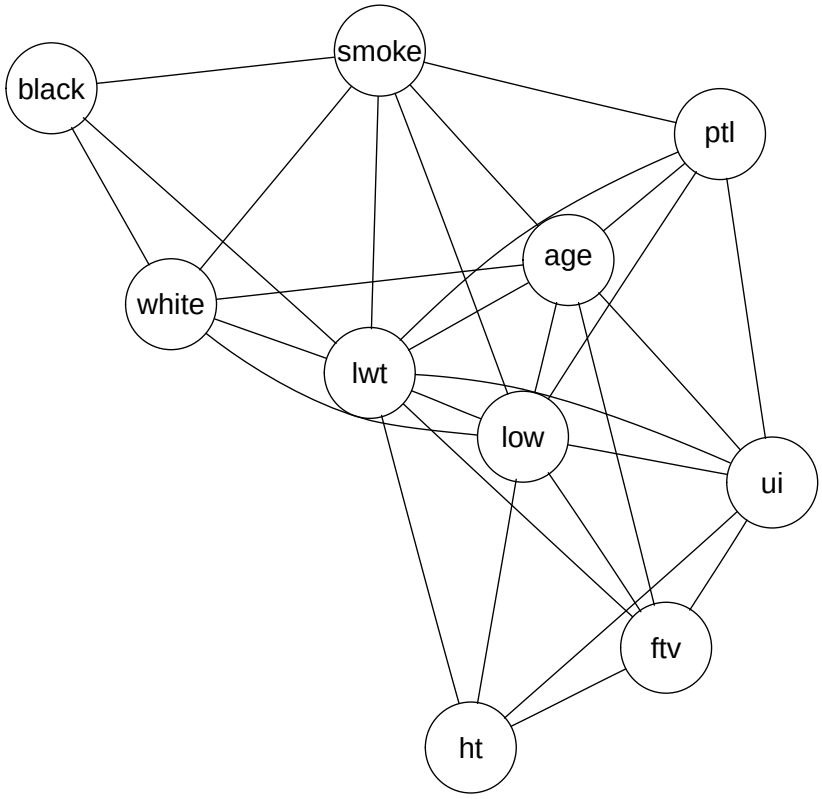
```
#Modello ottenuto con il criterio BIC
m1 <- fit.matrix(fit1)
rp1 <- fit.recall.precision(m1)

m1

##
##            predicted 0 predicted 1
##   actual 0         122           8
##   actual 1          43          16

rp1

##      recall precision
## 1 0.2711864 0.6666667

#Modello ottenuto con il criterio AIC
m2 <- fit.matrix(fit2)
rp2 <- fit.recall.precision(m2)

m2

##
##            predicted 0 predicted 1
##   actual 0         118          12
##   actual 1          38          21

rp2

##      recall precision
## 1 0.3559322 0.6363636
```

Come possiamo osservare, la *precisione* di questi due modelli è maggiore rispetto al nostro modello di partenza.
Per quanto riguarda il *recupero*, invece, esso è minore del valore trovato in precedenza nel primo modello, ma rimane invariato nel secondo. Come abbiamo già detto, visto il contesto di questo studio, il *recupero* è più importante da massimizzare rispetto alla *precisione*, inoltre l'utilizzo di meno variabili è un vantaggio, per cui il modello iniziale è da considerarsi più accurato.

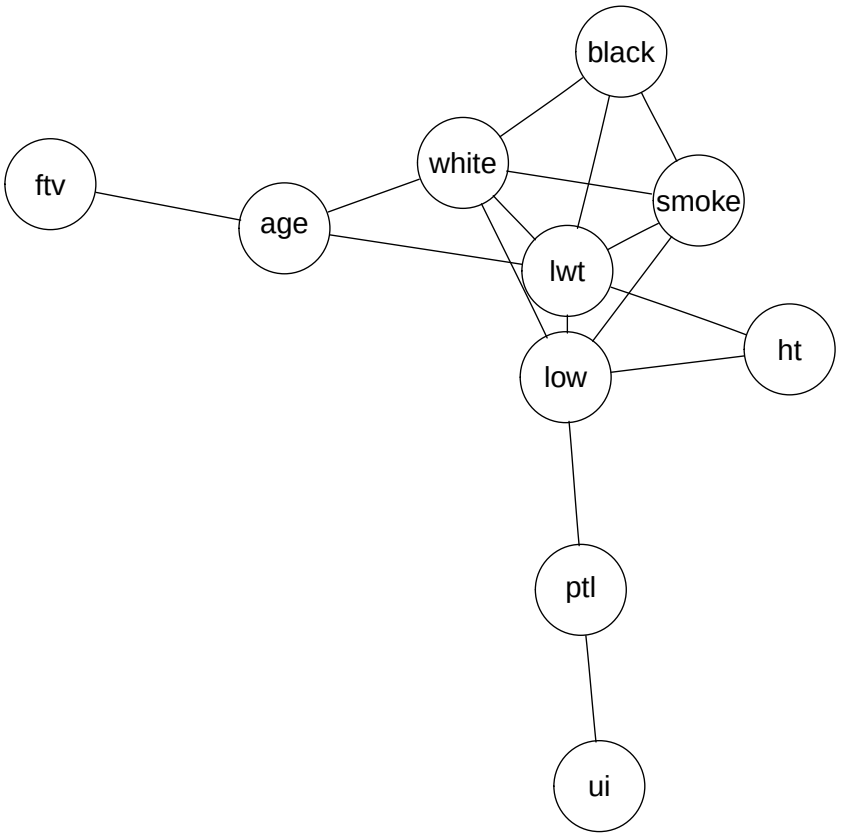
3.2 Grafi non orientati

Per quanto riguarda i grafi non orientati, si ha:

```
graph.LOW <- cmod(~.^., data = studyLOW)

#Grafo non orientato, AIC
aic.LOW <- gRbase::stepwise(graph.LOW)
plot(as(aic.LOW, "graphNEL"), "fdp")
```




```
#Grafo non orientato, BIC
bic.LOW <- gRbase::stepwise(graph.LOW,  k = log(nrow(studyBWT)))
plot(as(bic.LOW, "graphNEL"), "fdp")
```



Si nota che:

1. il numero di variabili da cui **low** è dipendente secondo il grafo generato dal criterio *AIC* è molto elevato: la variabile dipende da tutte le altre variabile tranne che da **black**;
2. osservando il grafo generato da *BIC*, invece, si osserva che si sostituisce la variabile **ui** con la variabile **ptl** nel modello che fino ad ora consideriamo il più accurato.

Calcoliamo i modelli:

```
attach(studyLOW)

fit3 <- glm(low ~ lwt + smoke + ht + ui + white + ptl + age + ftv,
            family = binomial)
summary(fit3)
```

```
##
## Call:
## glm(formula = low ~ lwt + smoke + ht + ui + white + ptl + age +
##     ftv, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8671  -0.8271  -0.5240   0.9804   2.2208
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.05954    0.27374  -3.871 0.000109 ***
## lwt         -0.42676    0.20104  -2.123 0.033772 *
## smoke        0.99338    0.39589   2.509 0.012100 *
## ht           1.84600    0.69611   2.652 0.008004 **
## ui           0.75137    0.46186   1.627 0.103777
## white       -1.01865    0.39742  -2.563 0.010373 *
## ptl          0.26546    0.17059   1.556 0.119676
## age         -0.16648    0.19547  -0.852 0.394406
## ftv          0.07401    0.18072   0.410 0.682161
## ---
## 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: 201.81  on 180  degrees of freedom
## AIC: 219.81
##
## Number of Fisher Scoring iterations: 4
```

```
fit4 <- glm(low ~ lwt + smoke + ht + ptl + white, family = binomial)
summary(fit4)
```

```
##
## Call:
## glm(formula = low ~ lwt + smoke + ht + ptl + white, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7335  -0.8358  -0.5436   0.9474   2.1421
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.9054     0.2582  -3.507 0.000453 ***
## lwt          -0.4662     0.1998  -2.334 0.019614 *
```

```
## smoke         0.9981     0.3890   2.566 0.010290 *
## ht            1.7368     0.6940   2.503 0.012325 *
## ptl           0.2936     0.1654   1.774 0.076009 .
## white        -1.0413     0.3862  -2.696 0.007008 **
## ---
## 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: 205.39  on 183  degrees of freedom
## AIC: 217.39
##
## Number of Fisher Scoring iterations: 4
```

Per quanto riguarda il secondo modello, si osserva che la significatività delle variabili è nettamente minore rispetto a quella delle variabili del nostro modello migliore. Il primo modello invece presenta una buona significatività delle variabili.

```
#Modello ottenuto dal grafo AIC
m3 <- fit.matrix(fit3)
rp3 <- fit.recall.precision(m3)

m3

##
##            predicted 0 predicted 1
##   actual 0         117          13
##   actual 1          37          22

rp3

##      recall precision
## 1 0.3728814 0.6285714

#Modello ottenuto dal grafo BIC
m4 <- fit.matrix(fit4)
rp4 <- fit.recall.precision(m4)

m4

##
##            predicted 0 predicted 1
##   actual 0         119          11
##   actual 1          40          19
```

```
rp4

##      recall precision
## 1 0.3220339 0.6333333
```

Si nota inoltre che il *recupero* del modello generato dal grafo di tipo *AIC* è maggiore (così come lo è anche la precisione).
Ciononostante, il modello prevede comunque un numero elevato di variabili non significative al suo interno per un aumento non elevato del *recupero*.
Un'ulteriore studio che possiamo fare per capire quale dei due modelli sia migliore è il *test del rapporto di verosimiglianza*:

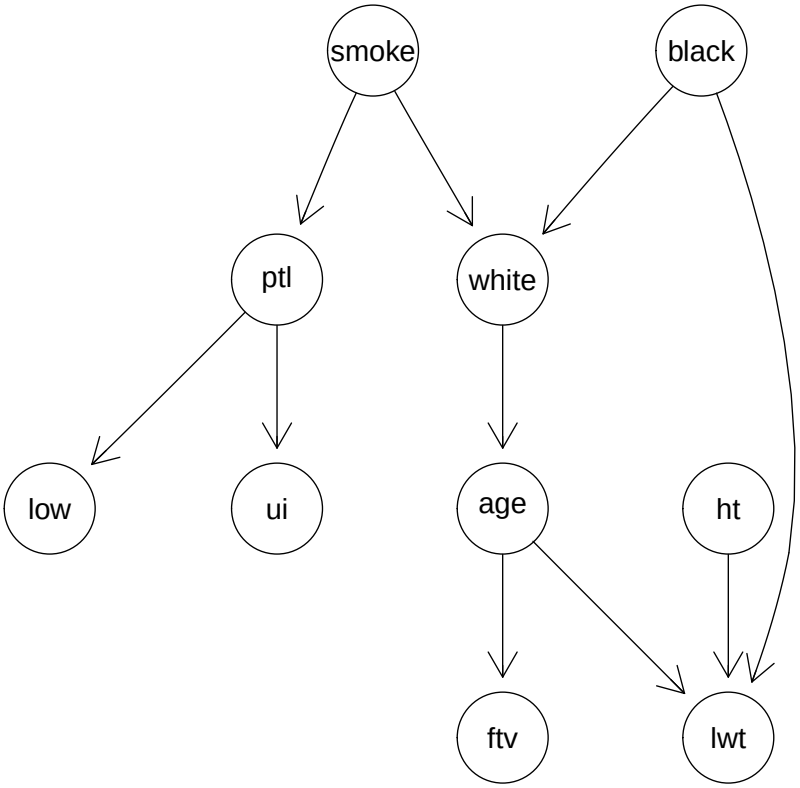
```
lrtest(fit, fit3)

## Likelihood ratio test
##
## Model 1: low ~ lwt + smoke + ht + ui + white
## Model 2: low ~ lwt + smoke + ht + ui + white + ptl + age + ftv
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -102.38
## 2   9 -100.91  3 2.9507     0.3993
```

Osservando questo test, diventa chiaro che il modello iniziale è ancora da considerarsi superiore.

3.2.1 DAG

Infine, osserviamo il DAG.

```
block_ <- c(3, 1, 2, 1, 2, 1, 2, 3, 1, 1)
blM_ <- matrix(0, nrow = 10, ncol = 10)
rownames(blM_) <- colnames(blM_) <- names(studyLOW)
for (b in 2:3) blM_[block_==b, block_<b] <- 1

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

dag1.LOW <- hc(studyLOW, blacklist = blackL_)
plot(as(amat(dag1.LOW), "graphNEL"))
```



Diversamente dalle nostre aspettative, il DAG ottenuto è completamente diverso da quello precedente.
Inoltre solo una variabile sembra responsabile della variazione della variabile obiettivo: la variabile **ptl**, che non influiva nello scorso studio.
Proviamo a costruire il modello con questa variabile:

```
attach(studyLOW)

fit5 <- glm(low ~ ptl)
summary(fit5)

##
## Call:
## glm(formula = low ~ ptl)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -0.8300  -0.2760  -0.2760   0.5393   0.7240
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.31217    0.03323   9.395  < 2e-16 ***
## ptl          0.09110    0.03332   2.735  0.00685 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2086718)
##
##     Null deviance: 40.582  on 188  degrees of freedom
## Residual deviance: 39.022  on 187  degrees of freedom
## AIC: 244.19
##
## Number of Fisher Scoring iterations: 2
```

E ora osserviamone i valori di *recupero* e *precisione*:

```
m5 <- fit.matrix(fit5)
rp5 <- fit.recall.precision(m5)

m5

##
##            predicted 0 predicted 1
##   actual 0         126           4
##   actual 1          57           2

rp5

##       recall precision
## 1 0.03389831 0.3333333
```

Entrambi i valori sono inferiori a quelli ottenuti nei modelli precedenti, quindi questo modello viene scartato.

3.3 Variabile lwt alla terza e non alla prima

Come ultimo passo nello studio di questo modello, si è deciso di considerare il risultato ottenuto alla fine dell'analisi del modello di regressione lineare multipla: cioè il fatto che la variabile **lwt** pare essere più rilevante se posta al cubo.

```
attach(studyLOW)

fit6 <- glm(low ~ I(lwt^3) + smoke + ht + ui + white, family = binomial)
summary(fit6)
```

```
##
## Call:
## glm(formula = low ~ I(lwt^3) + smoke + ht + ui + white, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6923  -0.8111  -0.4855   1.0553   2.1005
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.97092    0.26590  -3.651 0.000261 ***
## I(lwt^3)    -0.09745    0.06024  -1.618 0.105721
## smoke        1.14516    0.38454   2.978 0.002902 **
## ht           1.91844    0.73043   2.626 0.008628 **
## ui           0.95830    0.44366   2.160 0.030772 *
## white       -1.11663    0.38513  -2.899 0.003739 **
## ---
## 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: 204.23  on 183  degrees of freedom
## AIC: 216.23
##
## Number of Fisher Scoring iterations: 6

m6 <- fit.matrix(fit6)
rp6 <- fit.recall.precision(m6)

m6

##
##            predicted 0 predicted 1
##   actual 0         114          16
##   actual 1          37          22

rp6

##      recall precision
## 1 0.3728814 0.5789474
```

Come si può osservare, quindi, la variabile **low** è meno significativa nel modello se posta al cubo (a differenza di come accadeva in precedenza) ma il *recupero* aumenta, raggiungendo lo stesso del modello derivato dallo studio del grafo di criterio *BIC*. Possiamo concludere la nostra analisi, quindi, osservando che

questo ultimo modello permette di ricevere il *recupero* più alto trovato in questo studio utilizzando meno variabili dell'altro.