UNIVERSITÀ
DEGLI STUDI
FIRENZE

Corso di Laurea Magistrale
Data Science, Calcolo Scientifico e IA

Relazione di Fondamenti di Statistica per Data Science

MICHAEL CAVICCHIOLI

Anno Accademico 2023-2024

Michael Cavicchioli: Corso di Laurea Magistrale
Data Science, Calcolo Scientifico e IA, © Anno Accademico 2023-2024

INDICE

1

INTRODUZIONE

Il presente studio si concentra su un esperimento di intervento completamente randomizzato, mirato a valutare gli effetti di un programma su individui disoccupati ad alto rischio. L'obiettivo principale dell'intervento è prevenire la compromissione della salute mentale e promuovere il reinserimento lavorativo di alta qualità.

1.1 CONTESTO E DESIGN SPERIMENTALE

L'esperimento coinvolge due condizioni: il gruppo di controllo, che riceve un libretto contenente brevi descrizioni sui metodi di ricerca del lavoro e suggerimenti, e il gruppo di intervento, sottoposto a cinque sessioni attive di formazione. L'intervento è mirato specificamente a individui con un punteggio di rischio superiore a una soglia prestabilita (1.38), basato su tensioni finanziarie, assertività e punteggio di depressione.

1.2 VARIABILI DI STUDIO

Le variabili preso-trattamento includono informazioni demografiche come sesso, età, etnia e stato civile, insieme alla variabile "Risk" che indica se il partecipante è ad alto rischio o meno. Le variabili post-trattamento comprendono il livello di consumo di alcol dopo 6 mesi, la depressione misurata con una sottoscala di 11 elementi basata sulla Hopkins Symptom Checklist e la variabile binaria "employ6", che indica se il soggetto è impiegato per almeno 20 ore a settimana sei mesi dopo l'assegnazione all'intervento.

1.3 OBIETTIVI DELLO STUDIO

L'obiettivo principale dello studio è valutare l'impatto dell'intervento sul livello di depressione e sulla ricollocazione lavorativa.

1.4 METODOLOGIA STATISTICA

Per raggiungere questi obiettivi, saranno impiegate analisi statistiche come modelli di regressione e reti Bayesiane, al fine di identificare le relazioni significative tra le variabili di interesse.

2

IL CAMPIONE

2.1 CARICAMENTO LIBRERIE E DATASET

Prima di tutto, si sono caricati sia il *dataset*, che le librerie che si andranno ad utilizzare: il primo per ovvie ragione, mentre, le seconde, perchè forniscono metodi utili, già creati e testati, per operare con i dati.
Il codice **R** è stato il seguente:

```
# Carico le librerie che mi serviranno
library(foreign)
library(scales)
library(ggplot2)
library(gRbase)
library(gRain)
library(gRim)
library(bnlearn)
library(igraph)

# Carico il file contenente funzioni utili
source("C:\\Users\\Cavicchioli\\Desktop\\Cavicchioli_Michael_7149344\\utils
    .R")
# Carico il dataset
jobs <- read.table("C:\\Users\\Cavicchioli\\Desktop\\Cavicchioli_Michael_
    7149344\\JOBSII.txt", header=T)

# Sostituisco i valori NA con -1, per non avere problemi in seguito
jobs$D[is.na(jobs$D)] <- -1

# Estrapolo le colonne di interesse
job <- jobs[c(1,2,3,4,7:10, 18:20)]
```

Si è notato che tutti i valori NA di D corrispondono a valori 0 di Z, pertanto, per non incorrere in problemi di valori mancanti, usando metodi di alcune librerie, si è deciso di sostituire i valori NA, della variabile D, con -1, questo per avere la seguente casistica:

- −1: trattamento non proposto.
- 0: trattamento rifiutato.
- 1: trattamento accettato.

2.2 ANALISI PRELIMINARE

Prima di procedere con le analisi statistiche, è buona norma studiare e cercare di capire il tipo di campione (*dataset*) che si ha davanti, questo per evitare perdite di tempo sul fatto che qualcosa non torni e per farsi un'idea di come, eventualmente, procedere, e quali strategie attuare.

2.2.1 *Distribuzione dell'età*

Il seguente codice serve per generare un istogramma, raffigurante la distribuzione dell'età della popolazione all'interno del *dataset*.

```
# Distribuzione dell'eta'
hist(job$age, main="Distribuzione dell'eta'", xlab="Eta'", ylab="#.",
     xlim = c(0, 100), col="lightblue")
```

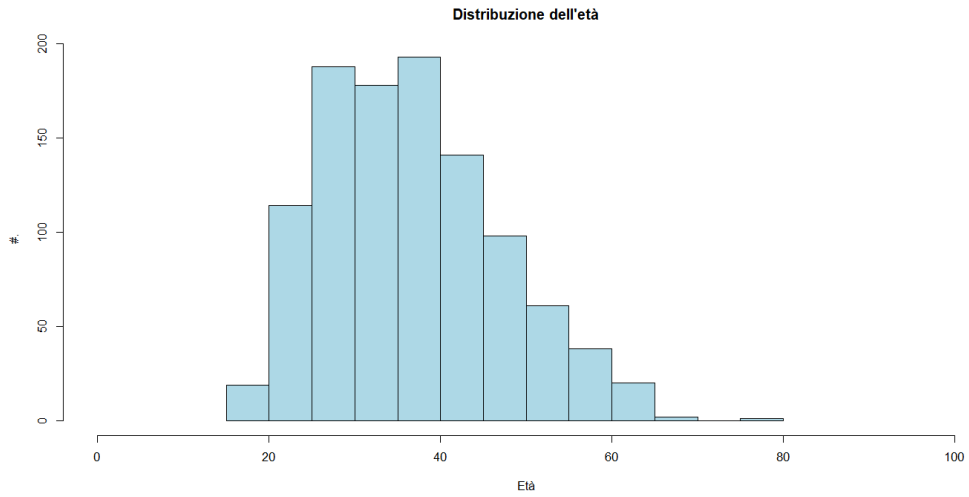

Figura 1: Distribuzione dell'età della popolazione presente nel *dataset*.

Com'è possibile dedurre dal grafico, la maggioranza della popolazione ha un'età compresa tra i 25 e i 40 anni, pochi casi di età minore ai 20 e rarissimi casi over 65.

2.2.2 *Distribuzione sesso ed etnia*

Dopodichè, si è andati a controllare la distribuzione del sesso e dell'etnia, mediante il codice proposto successivamente. La figura 2 rappresenta quanto descritto.

```
# Imposta la griglia dei grafici su una riga e due colonne
par(mfrow=c(1, 2))

# Grafico a barre per la variabile sesso
barplot(table(job$sex), main="Distribuzione per Sesso", xlab="Sesso",
        ylab="#.", names.arg = c("M", "F"), col="lightblue")

# Grafico a barre per la variabile etnia
barplot(table(job$race), main="Distribuzione per Etnia", xlab="Etnia",
        ylab="#.", names.arg = c("Bianca", "Altro"), col="lightgreen")

# Ripristina la griglia dei grafici
par(mfrow=c(1, 1))
```

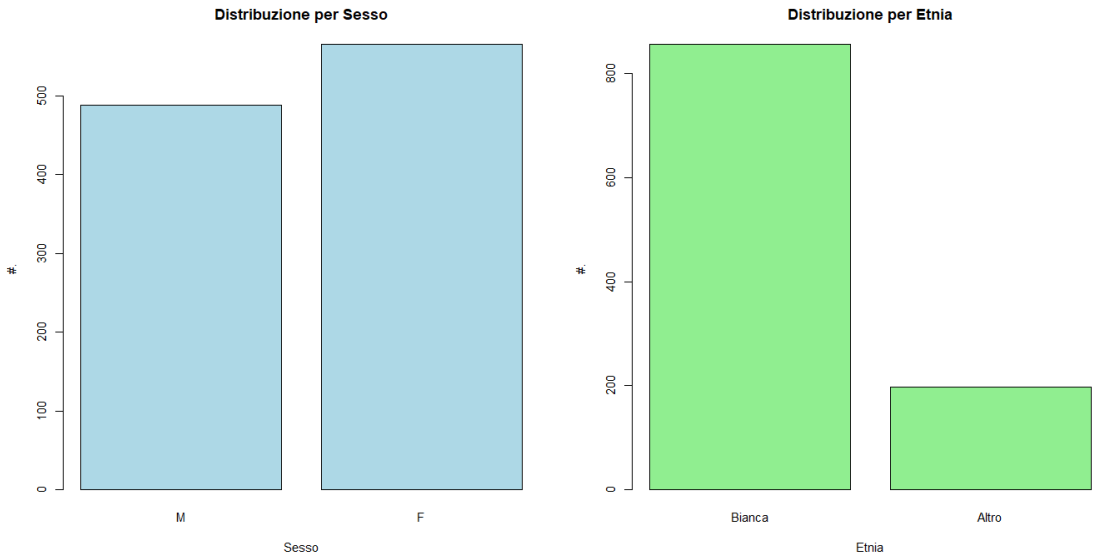

Figura 2: Distribuzione del sesso e dell'etnia presenti nel *dataset*.

Si osserva che la popolazione è per lo più femminile, mentre si ha una maggioranza dell'etnia bianca rispetto al resto.

2.2.3 *Distribuzione depressione per età*

Successivamente, si è creato uno Scatter plot in cui si è allocato le età dei soggetti del *dataset* sull'asse delle ascisse (x), mentre i molteplici gradi di depressione[1][2], passati 6 mesi, sull'asse delle ordinate (y).

```
# Distribuzione della depressione per eta'
plot(job$age, job$depress6, xlab = "Eta'", ylab = "Liv. Depr.",
    main = "Distribuzione della depressione per eta'",
    col = "Red", pch = 16, cex = 2)
```

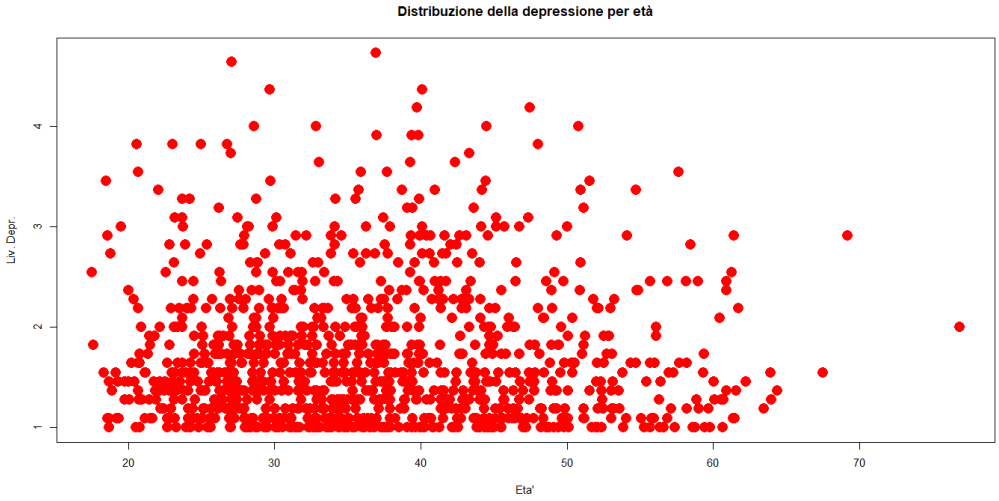

Figura 3: Distribuzione dei vari livelli di depressione per ogni persona.

Dal grafico si può evincere che si è in presenza di un *cluster* di persone, con età compresa tra i 25 e i 40 anni, che hanno il livello di depressione compreso tra 1 e 2, rispetto al resto della popolazione.

2.2.4 *Distribuzione trattamento per etnie*

Un'altra considerazione che si è fatta, è quella di rappresentare tramite un grafico a torta, la suddivisione della proposta, o meno, di trattamento, e la sua accettazione e rifiuto, per le due categorie di etnie presenti nel

1 Si fa notare che per il momento non si è ancora normalizzato il valore della depressione (0,1), perchè per questo tipo di grafico è risultato opportuno mantenere tutti i valori iniziali della variabile in questione.

2 Il grafico riporta tutti i soggetti presenti nel *dataset*, ovvero coloro che si sono rifiutati al trattamento, coloro che lo hanno accettato e chi non lo ha ricevuto.

dataset.
Per fare questo, si è utilizzato il seguente codice **R** e il risultato è mostrato in figura 4:

```r
# Estraggo i valori per le categorie specificate
valori <- c(sum(job$Z==1 & job$D==1 & job$race==1),
            sum(job$Z==1 & job$D==1 & job$race==0),
            sum(job$Z==1 & job$D==0 & job$race==1),
            sum(job$Z==1 & job$D==0 & job$race==0),
            sum(job$Z==1 & job$race==1),
            sum(job$Z==1 & job$race==0))

# Etichette per le fette
etichette <- c("Z=1, D=1, race=1",
            "Z=1, D=1, race=0",
            "Z=1, D=0, race=1",
            "Z=1, D=0, race=0",
            "Z=0, race=1",
            "Z=0, race=0")

percentuali <- percent(valori / sum(valori))

pie(valori, labels = percentuali, col = rainbow(length(valori)))
legend(x=1.2, y=0.8, horiz = FALSE, text.width = 0.4,
    legend = etichette, cex=0.6, fill = rainbow(length(valori)))

# Titolo
title("Distribuzione del trattamento per le etnie")
```

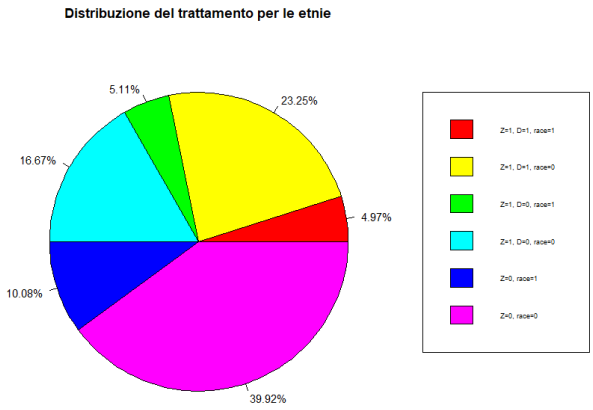

Figura 4: Distribuzione del trattamento per le etnie presenti nel *dataset*.

Dal grafico si può notare che il trattamento è stato somministrato esattamente a metà popolazione, però, solo circa il 28% di quest'ultimi ha aderito. Infatti, il 22% circa lo ha rifiutato, dove il 16.67% sono persone di

etnia bianca.
Fino ad ora, l'analisi è stata prettamente generale, facendo alcune osservazioni sul *dataset* senza andare nello specifico e senza osservare gli ipotetici benefici che il trattamento possa aver creato, sia nel livello depressivo, che in quello del reimpiego.
Pertanto, le successive considerazioni si baseranno sull'azione del trattamento fini agli scopi precedentemente citati.

2.2.5 *Interazione trattamento, depressione e rischio*

Proseguendo con l'analisi, si è andati ad osservare il livello di depressione, per le persone a cui è stato assegnato e non il trattamento, influenzato dal fattore rischio, cioè se i disoccupati sono ad alto rischio o meno.
Il codice che ha permesso ciò è il seguente.

```
# Interazione tra trattamento, depressione e rischio
dep6 <- as.numeric(job$depress6 <= 1.5)
ggplot(job, aes(x = Z, y = dep6, color = factor(Risk))) +
  geom_point(position = position_jitter(width = 0.2), size = 3) +
  labs(title = "Interazione tra Trattamento, Depressione e Rischio",
       x = "Trattamento (0 = non assegnato, 1 = assegnato)",
       y = "Depressione (0 = Alta, 1 = Bassa)",
       color = "Rischio (0=Basso, 1=Alto)") +
  scale_color_manual(values = c("blue", "red")) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5)
  )
```

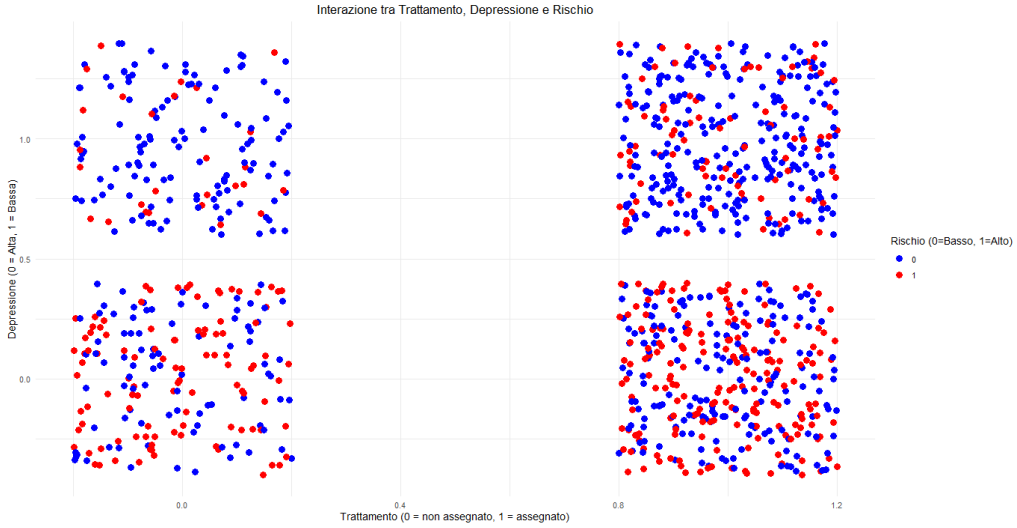

Figura 5: Livelli di depressione, a seconda del trattamento, influenzati dal fattore rischio.

Dal grafico di dispersione si può osservare che, per la popolazione a cui non è stato assegnato il trattamento, si hanno molti più casi di alta depressione, dovuta ad un elevato rischio, rispetto a coloro che hanno una depressione più lieve, ma con un rischio minore.
Mentre, per coloro a cui è stato assegnato il trattamento, si può osservare che ci sono molti più casi di depressione, sia lievi, che non, rispetto agli altri. Inoltre, sono presenti molti casi di alta depressione per coloro a cui è stato assegnato il trattamento, derivanti sia da un rischio elevato, che non.
Sembrerebbe, quindi, che il trattamento non abbia avuto efficacia, se non per il fatto che questo grafico non tiene conto di coloro che hanno rifiutato il trattamento, nonostante gli sia stato assegnato.

2.2.6 *Interazione trattamento, alcol e reimpiego*

Dopodichè, si è provato a fare la stessa osservazione della sezione precedente, ma tra il trattamento, alcol e il reimpiego.
Successivamente sono riportati sia il codice, che il grafico.

```
# Interazione tra trattamento, alcol e reimpiego
ggplot(job, aes(x = factor(Z), y = alcohol6, fill = factor(employ6))) +
  geom_boxplot() +
  labs(x = "Trattamento (0 = non assegnato, 1 = assegnato)",
       y = "Livelli di Alcol",
```

```
      fill = "Reimpiego dopo 6 mesi") +
scale_fill_manual(values = c("red", "blue"), breaks = c(0, 1), labels = c
  ("No", "Si")) +
theme_minimal()
```

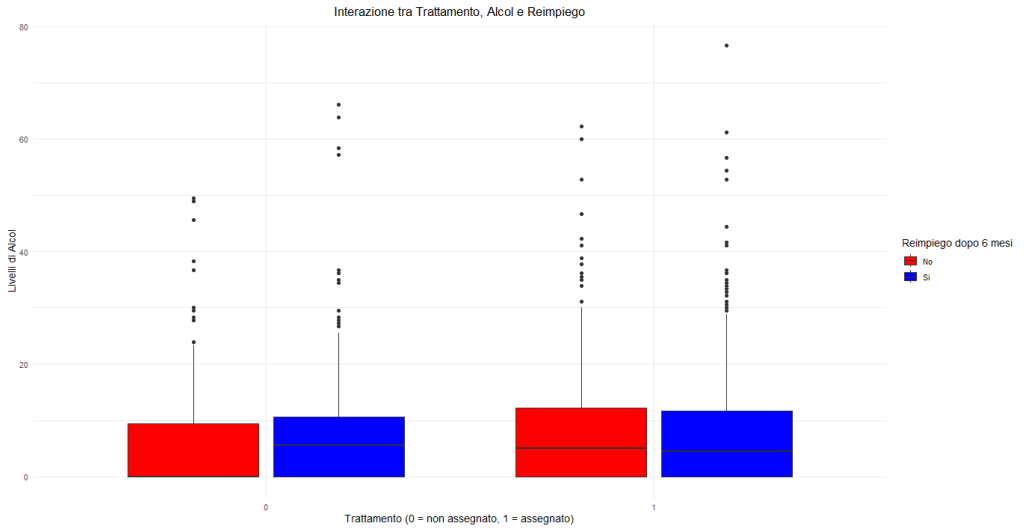

Figura 6: Interazione trattamento, alcol e reimpiego.

E' possibile notare che il comportamento per il campione dei trattati e dei non trattati è simile. Si vede che non vi è una netta differenza in termini di reimpiego lavorativi per i due sotto campioni. Risulta essere leggermente più alta la percentuale degli occupati per campione dei trattati rispetto a quello dei non trattati. Inoltre, per i trattati, dato che la mediana è posizionata più in basso rispetto al centro del box, è possibile affermare che la distribuzione dei livelli di alcol tende verso valori più bassi, quindi vi è un minor consumo.
Da questa analisi sembrerebbe che il trattamento non abbia avuto un grande effetto in termini di reimpiego, tuttavia è da valutare il comportamento nei modelli.

3

REGRESSIONE LOGISTICA

Noto come modello logit, modello logistico o regressione logistica, è un modello non lineare, il quale viene utilizzato quando si è in presenza di una variabile risposta (dipendente) di tipo dicotomico.
Pertanto, dato che i principali risultati di interesse sono il livello di depressione, 0 = alto e 1 = basso, e il reimpiego, anch'essa variabile binaria, ma con 1 = se il soggetto lavora più di 20 ore a settimana e 0 altrimenti, si è scelto di utilizzare il modello sopra descritto, andando a sostituire i valori di depressione presenti con quelli di interesse, impostando 1 se il livello di depressione nel *dataset* era minore o uguale di 1.5, 0 altrimenti.

3.1 CRITERI DI SELEZIONE DEL MODELLO E METODI DI PENALIZZAZIONE

Per il problema attuale, si è scelto di utilizzare i criteri di *forward*, *backward* e *both*, per vedere quale metodo risultasse il migliore per la scelta del modello.
Il loro funzionamento è spiegato brevemente:

- *Forward*: inizia con un modello vuoto e aggiunge progressivamente le variabili più significative o informative, una alla volta, fino a quando un certo criterio di arresto viene soddisfatto (ad esempio, un miglioramento minimo della performance del modello).
- *Backward*: inizia con un modello contenente tutte le variabili e rimuove iterativamente quelle meno significative, di solito basandosi su test di significatività statistica o altri criteri definiti a priori.
- *Both*: combina elementi di *forward* e *backward* selection. Inizia con un modello vuoto, aggiunge o rimuove variabili in base a criteri specifici ad ogni passo, valutando costantemente l'impatto sulla

performance del modello. Questo approccio può includere anche l'aggiunta e la rimozione di più variabili contemporaneamente.

In aggiunta, si è scelto di utilizzare, per ogni criterio, sia il metodo di penalizzazione *AIC*, che *BIC*, uno alla volta. Il loro funzionamento è brevemente spiegato qui:

- *AIC*: penalizza modelli meno complessi. L'obiettivo è trovare il modello che minimizza l'*AIC*, bilanciando bene adattamento e complessità.
- *BIC*: basato su principi Bayesiani, penalizza modelli più complessi, scegliendo modelli più parsimoniosi (meno complessi).

Si è usato anche il metodo della *MV*, ovvero quel metodo che si basa sul selezionare la combinazione di variabili che massimizza la verosimiglianza del modello, il quale non è un metodo di penalizzazione.

3.2 IL REIMPIEGO

In questa sezione, come prima fase della regressione logistica, si è studiato il reimpiego delle persone.
Il primo passo è stato quello di 'normalizzare' il *dataset*, andando ad impostare come variabili binarie le variabili raffiguranti i livelli di alcol e di depressione, dove entrambe indicano con 0 un alto livello e con 1 il viceversa. Questa normalizzazione ha effetti permanenti, pertanto non sono riportati nella sezione successiva.
In seguito, si sono creati, e allenati, due modelli, i quali hanno come variabile dipendente il reimpiego, dove il primo è composto da tutte le variabili del *dataset* (pieno), mentre il secondo è il modello vuoto con solo l'intercetta, per poi osservare i risultati a schermo.
Il codice è stato il seguente.

```
# Reimpiego dopo 6 mesi
job$alcohol6 = as.numeric(job$alcohol6<=4.5)
job$depress6 = as.numeric(job$depress6<=1.5)

fit <- glm(job$employ6 ~ job$Risk + job$sex + job$age + job$race + job$Z
    + job$D + job$nonmarried + job$alcohol6 + job$depress6,
    family = binomial)
summary(fit)

Call:
glm(formula = job$employ6 ~ job$Risk + job$sex + job$age + job$race +
```

```
    job$Z + job$D + job$nonmarried + job$alcohol6 + job$depress6,
    family = binomial)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.639998   0.381848   4.295 1.75e-05 ***
job$Risk        0.043564   0.141969   0.307   0.7590
job$sex        -0.284226   0.135946  -2.091   0.0366 *
job$age        -0.029692   0.006771  -4.385 1.16e-05 ***
job$race       -0.663253   0.167268  -3.965 7.33e-05 ***
job$Z          -0.256689   0.290930  -0.882   0.3776
job$D           0.114762   0.163612   0.701   0.4830
job$nonmarried  0.218007   0.139493   1.563   0.1181
job$alcohol6    0.040253   0.135982   0.296   0.7672
job$depress6    0.570937   0.139932   4.080 4.50e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1396.4  on 1052  degrees of freedom
Residual deviance: 1334.8  on 1043  degrees of freedom
AIC: 1354.8

Number of Fisher Scoring iterations: 4
```

```
fit_intercetta <- glm(job$employ6 ~ 1, family = binomial)
summary(fit_intercetta)

Call:
glm(formula = job$employ6 ~ 1, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.49818    0.06356   7.839 4.56e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1396.4  on 1052  degrees of freedom
Residual deviance: 1396.4  on 1052  degrees of freedom
AIC: 1398.4

Number of Fisher Scoring iterations: 4
```

E' possibile osservare dall'allenamento del modello completo, che le

variabili Age, Race e Depress6 hanno un alto impatto sulla variabile dipendente, mentre Sex ne ha poca e le altre per niente.
Successivamente, si è andati ad applicare tutti i criteri di selezione del modello e metodi di penalizzazione, spiegati in 3.1.
Di seguito sono riportati i codici utilizzati e le relative informazioni ricavate.

```
# BACKWARD - MV
back_mv <- step(fit, scope = formula(fit_intercetta), direction = "backward
    ", k = 0)
summary(back_mv)

Call:
glm(formula = job$employ6 ~ job$Risk + job$sex + job$age + job$race +
    job$Z + job$D + job$nonmarried + job$alcohol6 + job$depress6,
    family = binomial)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.639998   0.381848   4.295 1.75e-05 ***
job$Risk        0.043564   0.141969   0.307   0.7590
job$sex        -0.284226   0.135946  -2.091   0.0366 *
job$age        -0.029692   0.006771  -4.385 1.16e-05 ***
job$race       -0.663253   0.167268  -3.965 7.33e-05 ***
job$Z          -0.256689   0.290930  -0.882   0.3776
job$D           0.114762   0.163612   0.701   0.4830
job$nonmarried  0.218007   0.139493   1.563   0.1181
job$alcohol6    0.040253   0.135982   0.296   0.7672
job$depress6    0.570937   0.139932   4.080 4.50e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1396.4  on 1052  degrees of freedom
Residual deviance: 1334.8  on 1043  degrees of freedom
AIC: 1354.8

Number of Fisher Scoring iterations: 4

print(BIC(back_mv))
[1] 1404.346
```

```
# BACKWARD - AIC
back_aic <- step(fit, scope = formula(fit_intercetta), direction = "
    backward", k = 2)
summary(back_aic)
```

```
Call:
glm(formula = job$employ6 ~ job$sex + job$age + job$race + job$nonmarried +
    job$depress6, family = binomial)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.483575   0.298756   4.966 6.84e-07 ***
job$sex        -0.277829   0.132871  -2.091   0.0365 *
job$age        -0.028928   0.006676  -4.333 1.47e-05 ***
job$race       -0.666482   0.164029  -4.063 4.84e-05 ***
job$nonmarried  0.218372   0.138070   1.582   0.1137
job$depress6    0.560656   0.132771   4.223 2.41e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1396.4  on 1052  degrees of freedom
Residual deviance: 1335.8  on 1047  degrees of freedom
AIC: 1347.8

Number of Fisher Scoring iterations: 4

print(BIC(back_aic))
[1] 1377.518
```

```
# BACKWARD - BIC
back_bic <- step(fit, scope = formula(fit_intercetta), direction = "
    backward", k = log(length(job$employ6)))
summary(back_bic)

Call:
glm(formula = job$employ6 ~ job$age + job$race + job$depress6,
    family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.561851   0.256894   6.080 1.20e-09 ***
job$age      -0.031909   0.006412  -4.977 6.47e-07 ***
job$race     -0.666585   0.163274  -4.083 4.45e-05 ***
job$depress6  0.555710   0.132070   4.208 2.58e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 1396.4  on 1052  degrees of freedom
Residual deviance: 1341.9  on 1049  degrees of freedom
AIC: 1349.9

Number of Fisher Scoring iterations: 4

print(BIC(back_bic))
[1] 1369.74
```

```
# FORWARD - MV
forw_mv <- step(fit_intercetta, scope = formula(fit), direction = "forward"
    , k = 0)
summary(forw_mv)

Call:
glm(formula = job$employ6 ~ job$age + job$depress6 + job$race +
    job$sex + job$nonmarried + job$Z + job$D + job$Risk + job$alcohol6,
    family = binomial)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.639998   0.381848   4.295 1.75e-05 ***
job$age        -0.029692   0.006771  -4.385 1.16e-05 ***
job$depress6    0.570937   0.139932   4.080 4.50e-05 ***
job$race       -0.663253   0.167268  -3.965 7.33e-05 ***
job$sex        -0.284226   0.135946  -2.091   0.0366 *
job$nonmarried  0.218007   0.139493   1.563   0.1181
job$Z          -0.256689   0.290930  -0.882   0.3776
job$D           0.114762   0.163612   0.701   0.4830
job$Risk        0.043564   0.141969   0.307   0.7590
job$alcohol6    0.040253   0.135982   0.296   0.7672
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1396.4  on 1052  degrees of freedom
Residual deviance: 1334.8  on 1043  degrees of freedom
AIC: 1354.8

Number of Fisher Scoring iterations: 4

print(BIC(forw_mv))
[1] 1404.346
```

```
# FORWARD - AIC
forw_aic <- step(fit_intercetta, scope = formula(fit), direction = "forward
```

```
    ", k = 2)
summary(forw_aic)

Call:
glm(formula = job$employ6 ~ job$age + job$depress6 + job$race +
    job$sex + job$nonmarried, family = binomial)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.483575   0.298756   4.966 6.84e-07 ***
job$age        -0.028928   0.006676  -4.333 1.47e-05 ***
job$depress6    0.560656   0.132771   4.223 2.41e-05 ***
job$race       -0.666482   0.164029  -4.063 4.84e-05 ***
job$sex        -0.277829   0.132871  -2.091   0.0365 *
job$nonmarried  0.218372   0.138070   1.582   0.1137
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1396.4  on 1052  degrees of freedom
Residual deviance: 1335.8  on 1047  degrees of freedom
AIC: 1347.8

Number of Fisher Scoring iterations: 4

print(BIC(forw_aic))
[1] 1377.518
```

```
# FORWARD - BIC
forw_bic <- step(fit_intercetta, scope = formula(fit), direction = "forward
    ", k = log(length(job$employ6)))
summary(forw_bic)

Call:
glm(formula = job$employ6 ~ job$age + job$depress6 + job$race,
    family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.561851   0.256894   6.080 1.20e-09 ***
job$age      -0.031909   0.006412  -4.977 6.47e-07 ***
job$depress6  0.555710   0.132070   4.208 2.58e-05 ***
job$race     -0.666585   0.163274  -4.083 4.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1396.4  on 1052  degrees of freedom
Residual deviance: 1341.9  on 1049  degrees of freedom
AIC: 1349.9

Number of Fisher Scoring iterations: 4

print(BIC(forw_bic))
[1] 1369.74
```

```
# BOTH - MV
both_mv <- step(fit, scope = formula(fit), direction = "both", k = 0)
summary(both_mv)

Call:
glm(formula = job$employ6 ~ job$Risk + job$sex + job$age + job$race +
    job$Z + job$D + job$nonmarried + job$alcohol6 + job$depress6,
    family = binomial)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.639998   0.381848   4.295 1.75e-05 ***
job$Risk        0.043564   0.141969   0.307   0.7590
job$sex        -0.284226   0.135946  -2.091   0.0366 *
job$age        -0.029692   0.006771  -4.385 1.16e-05 ***
job$race       -0.663253   0.167268  -3.965 7.33e-05 ***
job$Z          -0.256689   0.290930  -0.882   0.3776
job$D           0.114762   0.163612   0.701   0.4830
job$nonmarried  0.218007   0.139493   1.563   0.1181
job$alcohol6    0.040253   0.135982   0.296   0.7672
job$depress6    0.570937   0.139932   4.080 4.50e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1396.4  on 1052  degrees of freedom
Residual deviance: 1334.8  on 1043  degrees of freedom
AIC: 1354.8

Number of Fisher Scoring iterations: 4

print(BIC(both_mv))
[1] 1404.346
```

```
# BOTH - AIC
```

```
both_aic <- step(fit, scope = formula(fit), direction = "both", k = 2)
summary(both_aic)

Call:
glm(formula = job$employ6 ~ job$sex + job$age + job$race + job$nonmarried +
    job$depress6, family = binomial)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.483575   0.298756   4.966 6.84e-07 ***
job$sex        -0.277829   0.132871  -2.091   0.0365 *
job$age        -0.028928   0.006676  -4.333 1.47e-05 ***
job$race       -0.666482   0.164029  -4.063 4.84e-05 ***
job$nonmarried  0.218372   0.138070   1.582   0.1137
job$depress6    0.560656   0.132771   4.223 2.41e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1396.4  on 1052  degrees of freedom
Residual deviance: 1335.8  on 1047  degrees of freedom
AIC: 1347.8

Number of Fisher Scoring iterations: 4

print(BIC(both_aic))
[1] 1377.518
```

```
# BOTH - BIC
both_bic <- step(fit, scope = formula(fit), direction = "both", k = log(
    length(job$employ6)))
summary(both_bic)

Call:
glm(formula = job$employ6 ~ job$age + job$depress6 + job$race,
    family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.561851   0.256894   6.080 1.20e-09 ***
job$age      -0.031909   0.006412  -4.977 6.47e-07 ***
job$depress6  0.555710   0.132070   4.208 2.58e-05 ***
job$race     -0.666585   0.163274  -4.083 4.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1396.4  on 1052  degrees of freedom
Residual deviance: 1341.9  on 1049  degrees of freedom
AIC: 1349.9

Number of Fisher Scoring iterations: 4

print(BIC(both_bic))
[1] 1369.74
```

Com'è possibile osservare da questi risultati, si hanno valori più alti sia di *AIC*, che di *BIC*, per tutti quei modelli a cui è stata applicata la funzione di verosimiglianza come metodo di penalizzazione. Inoltre, riportando sempre il modello completo, essi includono anche variabili esplicative il cui significato è nullo.
Per quanto riguarda i modelli a cui è stato applicato il metodo di penalizzazione *AIC*, è possibile osservare che includono quasi tutte variabili esplicative fortemente significative, tranne il sesso, poco significativa e l'essere non sposati, per niente significativa. Inoltre, presentano un livello di *AIC* inferiore a tutti quanti, ma un livello di *BIC* maggiore rispetto ai modelli a cui è stato applicato quest'ultimo metodo di penalizzazione.
Infine, i restanti modelli, sottoposti a *BIC*, presentano solo variabili fortemente significative, un livello di *AIC* leggermente più alto di coloro sottoposti a questa penalizzazione (*AIC*), ma un livello inferiore di *BIC*.
Pertanto, si è preferito scegliere un modello ottenuto con questa tipologia di penalizzazione, ovvero il **back_bic**.

3.2.1 *Interpretazione coefficienti per il reimpiego*

I coefficienti ottenuti per il modello **back_bic** sono i seguenti.

```
# Coefficienti BACK_BIC
print(coefficients(back_bic))

 (Intercept)      job$age     job$race job$depress6
  1.56185149  -0.03190873  -0.66658473   0.55570991
```

Si nota che le variabili esplicative Age e Race hanno valori negativi, mentre Depress6 positivi.
Il fatto che ci siano valori negativi indica che queste variabili influenzano negativamente il reimpiego; mentre, la depressione, dato che ha un valore positivo, influenza positivamente l'essere reimpiegati.

Infine, si osserva che non è presente alcuna variabile inerente al trattamento: questo può voler dire che, probabilmente, quest'ultimo, non ha rilevanza significativa.

3.3 LA DEPRESSIONE

Anche in questo caso, si è adottato lo stesso *modus operandis* di quello in 3.2, senza però normalizzare il *dataset*.
Pertanto, in primis si è allenato il modello completo, dove l'unica variabile dipendente è stata Depress6, e successivamente quello vuoto, ovvero solo con l'intercetta.
Il codice è stato il seguente.

```
# Depressione dopo 6 mesi
fit <- glm(job$depress6 ~ job$Risk + job$sex + job$age + job$race + job$Z +
     job$D
           + job$nonmarried + job$alcohol6 + job$employ6, family = binomial
    )
summary(fit)

Call:
glm(formula = job$depress6 ~ job$Risk + job$sex + job$age + job$race +
    job$Z + job$D + job$nonmarried + job$alcohol6 + job$employ6,
    family = binomial)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.075757   0.393361   0.193   0.8473
job$Risk       -1.325336   0.139831  -9.478  < 2e-16 ***
job$sex         0.004544   0.135912   0.033   0.9733
job$age         0.003843   0.006851   0.561   0.5748
job$race        0.037850   0.173841   0.218   0.8276
job$Z          -0.294307   0.292909  -1.005   0.3150
job$D           0.335200   0.163981   2.044   0.0409 *
job$nonmarried -0.145622   0.140535  -1.036   0.3001
job$alcohol6    0.056376   0.136570   0.413   0.6798
job$employ6     0.571108   0.139783   4.086 4.39e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1455.0  on 1052  degrees of freedom
Residual deviance: 1328.5  on 1043  degrees of freedom
AIC: 1348.5
```

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

```
fit_intercetta <- glm(job$depress6 ~ 1, family = binomial)
summary(fit_intercetta)

Call:
glm(formula = job$depress6 ~ 1, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.13506    0.06177  -2.186   0.0288 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1455  on 1052  degrees of freedom
Residual deviance: 1455  on 1052  degrees of freedom
AIC: 1457

Number of Fisher Scoring iterations: 3
```

Dai risultati si può osservare che, da un primo allenamento, le variabili Risk ed Employ sono fortemente significative per Depress6, D (trattamento accettato/rifiutato) è poco significativa, mentre le altre per niente.
Successivamente, si è cercato di capire quale fosse il modello migliore utilizzando gli stessi criteri e metodi di quelli in 3.2.
Di seguito sono riportati i codici.

```
# BACKWARD - MV
back_mv <- step(fit, scope = formula(fit_intercetta), direction = "backward
    ", k = 0)
summary(back_mv)

Call:
glm(formula = job$depress6 ~ job$Risk + job$sex + job$age + job$race +
    job$Z + job$D + job$nonmarried + job$alcohol6 + job$employ6,
    family = binomial)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.075757   0.393361   0.193   0.8473
job$Risk       -1.325336   0.139831  -9.478  < 2e-16 ***
job$sex         0.004544   0.135912   0.033   0.9733
job$age         0.003843   0.006851   0.561   0.5748
job$race        0.037850   0.173841   0.218   0.8276
```

```
job$Z          -0.294307   0.292909  -1.005   0.3150
job$D           0.335200   0.163981   2.044   0.0409 *
job$nonmarried -0.145622   0.140535  -1.036   0.3001
job$alcohol6    0.056376   0.136570   0.413   0.6798
job$employ6     0.571108   0.139783   4.086 4.39e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1455.0  on 1052  degrees of freedom
Residual deviance: 1328.5  on 1043  degrees of freedom
AIC: 1348.5

Number of Fisher Scoring iterations: 4

print(BIC(back_mv))
[1] 1398.087
```

```
# BACKWARD - AIC
back_aic <- step(fit, scope = formula(fit_intercetta), direction = "
    backward", k = 2)
summary(back_aic)

Call:
glm(formula = job$depress6 ~ job$Risk + job$D + job$employ6,
    family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.005049   0.116603   0.043   0.9655
job$Risk    -1.335575   0.138253  -9.660  < 2e-16 ***
job$D        0.196596   0.078486   2.505   0.0122 *
job$employ6  0.544729   0.136313   3.996 6.44e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1455  on 1052  degrees of freedom
Residual deviance: 1332  on 1049  degrees of freedom
AIC: 1340

Number of Fisher Scoring iterations: 4

print(BIC(back_aic))
[1] 1359.848
```

```
# BACKWARD - BIC
back_bic <- step(fit, scope = formula(fit_intercetta), direction = "
    backward", k = log(length(job$depress6)))
summary(back_bic)

Call:
glm(formula = job$depress6 ~ job$Risk + job$employ6, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.02105    0.11599   0.181    0.856
job$Risk    -1.32948    0.13767  -9.657  < 2e-16 ***
job$employ6  0.53739    0.13570   3.960 7.49e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1455.0  on 1052  degrees of freedom
Residual deviance: 1338.3  on 1050  degrees of freedom
AIC: 1344.3

Number of Fisher Scoring iterations: 4

print(BIC(back_bic))
[1] 1359.198
```

```
# FORWARD - MV
forw_mv <- step(fit_intercetta, scope = formula(fit), direction = "forward"
    , k = 0)
summary(forw_mv)

Call:
glm(formula = job$depress6 ~ job$Risk + job$employ6 + job$D +
    job$nonmarried + job$Z + job$age + job$alcohol6 + job$race +
    job$sex, family = binomial)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.075757   0.393361   0.193   0.8473
job$Risk       -1.325336   0.139831  -9.478  < 2e-16 ***
job$employ6     0.571108   0.139783   4.086 4.39e-05 ***
job$D           0.335200   0.163981   2.044   0.0409 *
job$nonmarried -0.145622   0.140535  -1.036   0.3001
job$Z          -0.294307   0.292909  -1.005   0.3150
job$age         0.003843   0.006851   0.561   0.5748
job$alcohol6    0.056376   0.136570   0.413   0.6798
```

```
job$race        0.037850   0.173841   0.218   0.8276
job$sex         0.004544   0.135912   0.033   0.9733
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1455.0  on 1052  degrees of freedom
Residual deviance: 1328.5  on 1043  degrees of freedom
AIC: 1348.5

Number of Fisher Scoring iterations: 4

print(BIC(forw_mv))
[1] 1398.087
```

```
# FORWARD - AIC
forw_aic <- step(fit_intercetta, scope = formula(fit), direction = "forward
    ", k = 2)
summary(forw_aic)

Call:
glm(formula = job$depress6 ~ job$Risk + job$employ6 + job$D,
    family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.005049   0.116603   0.043   0.9655
job$Risk    -1.335575   0.138253  -9.660  < 2e-16 ***
job$employ6  0.544729   0.136313   3.996 6.44e-05 ***
job$D        0.196596   0.078486   2.505   0.0122 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1455  on 1052  degrees of freedom
Residual deviance: 1332  on 1049  degrees of freedom
AIC: 1340

Number of Fisher Scoring iterations: 4

print(BIC(forw_aic))
[1] 1359.848
```

```
# FORWARD - BIC
forw_bic <- step(fit_intercetta, scope = formula(fit), direction = "forward
```

```
    ", k = log(length(job$depress6)))
summary(forw_bic)

Call:
glm(formula = job$depress6 ~ job$Risk + job$employ6, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.02105    0.11599   0.181    0.856
job$Risk    -1.32948    0.13767  -9.657  < 2e-16 ***
job$employ6  0.53739    0.13570   3.960 7.49e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1455.0  on 1052  degrees of freedom
Residual deviance: 1338.3  on 1050  degrees of freedom
AIC: 1344.3

Number of Fisher Scoring iterations: 4

print(BIC(forw_bic))
[1] 1359.198
```

```
# BOTH - MV
both_mv <- step(fit, scope = formula(fit), direction = "both", k = 0)
summary(both_mv)

Call:
glm(formula = job$depress6 ~ job$Risk + job$sex + job$age + job$race +
    job$Z + job$D + job$nonmarried + job$alcohol6 + job$employ6,
    family = binomial)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.075757   0.393361   0.193   0.8473
job$Risk       -1.325336   0.139831  -9.478  < 2e-16 ***
job$sex         0.004544   0.135912   0.033   0.9733
job$age         0.003843   0.006851   0.561   0.5748
job$race        0.037850   0.173841   0.218   0.8276
job$Z          -0.294307   0.292909  -1.005   0.3150
job$D           0.335200   0.163981   2.044   0.0409 *
job$nonmarried -0.145622   0.140535  -1.036   0.3001
job$alcohol6    0.056376   0.136570   0.413   0.6798
job$employ6     0.571108   0.139783   4.086 4.39e-05 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1455.0  on 1052  degrees of freedom
Residual deviance: 1328.5  on 1043  degrees of freedom
AIC: 1348.5

Number of Fisher Scoring iterations: 4

print(BIC(both_mv))
[1] 1398.087
```

```
# BOTH - AIC
both_aic <- step(fit, scope = formula(fit), direction = "both", k = 2)
summary(both_aic)

Call:
glm(formula = job$depress6 ~ job$Risk + job$D + job$employ6,
    family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.005049   0.116603   0.043   0.9655
job$Risk    -1.335575   0.138253  -9.660  < 2e-16 ***
job$D        0.196596   0.078486   2.505   0.0122 *
job$employ6  0.544729   0.136313   3.996 6.44e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1455  on 1052  degrees of freedom
Residual deviance: 1332  on 1049  degrees of freedom
AIC: 1340

Number of Fisher Scoring iterations: 4

print(BIC(both_aic))
[1] 1359.848
```

```
# BOTH - BIC
both_bic <- step(fit, scope = formula(fit), direction = "both", k = log(
    length(job$depress6)))
summary(both_bic)

Call:
```

```
glm(formula = job$depress6 ~ job$Risk + job$employ6, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.02105    0.11599   0.181    0.856
job$Risk    -1.32948    0.13767  -9.657  < 2e-16 ***
job$employ6  0.53739    0.13570   3.960 7.49e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1455.0  on 1052  degrees of freedom
Residual deviance: 1338.3  on 1050  degrees of freedom
AIC: 1344.3

Number of Fisher Scoring iterations: 4

print(BIC(both_bic))
[1] 1359.198
```

Anche in questo caso, valgono le stesse considerazioni fatte in 3.2, infatti: con il metodo di penalizzazione *AIC* si ottengono valori migliori di *AIC*, valori leggermente peggiori di *BIC* rispetto all'uso del medesimo metodo ed un modello con variabili fortemente significative, ma anche con una variabile poco significativa.
L'uso della funzione di verosimiglianza, anche questa volta, ha rappresentato il modello peggiore, sia in termini di variabile coinvolte (anche perchè include sempre il modello completo), che di livelli di *AIC* e *BIC*.
Pertanto, modelli ottenuti tramite il metodo *BIC* risultano essere più appropriati, perciò, anche questa volta, si è scelto il modello **back_bic**.

3.3.1 *Interpretazione coefficienti per la depressione*

I coefficienti ottenuti per il modello **back_bic** scelto sono i seguenti.

```
# Coefficienti BACK_BIC
print(coefficients(back_bic))

(Intercept)    job$Risk job$employ6
 0.02105079 -1.32947752  0.53738531
```

Il coefficiente negativo di Risk suggerisce che la sua presenza influisce negativamente sulla depressione, mentre quello di Employ6 positivamente, poichè positivo.

Inoltre, anche qui, non sono presenti alcune variabili inerenti al trattamento. Pertanto, vige la stessa considerazione fatta in 3.2.1, ovvero che, probabilmente, il trattamento è irrilevante.

4

RETI BAYESIANE E DAG

In questo capitolo vengono affrontati i DAG e le reti bayesiane per lo studio delle regressioni logistiche, dato che i DAG possono essere visti come una sequenza di tali regressioni.
Una rete bayesiana è conosciuta come modello grafico basato su un DAG. Il termine bayesiana si riferisce al fatto che viene utilizzata la medesima formula per un rapido calcolo delle probabilità aggiornate in sistemi complessi sfruttando la fattorizzazione fornita da un modello DAG.
I DAG, grafi aciclici diretti, sono particolari tipi di grafi in cui non esistono cicli diretti, ovvero che, partendo da un qualsiasi vertice del grafo e percorrendo tutti gli archi del grafo, non è possibile ritornare al vertice di partenza.

4.1 RIMOZIONE VARIABILI DAL DATASET

Innanzitutto si è rimossa la variabile ID, e quindi la colonna con tutti i valori annessi, dal *dataset*, questo poichè si è ritenuta essere una variabile per nulla significativa.
Per fare questo, si è utilizzato il seguente codice.

```
# Rimozione colonna ID
tutte_colonne <- colnames(job)
colonna_da_escludere <- c("ID")
colonne_da_mantenere <- setdiff(tutte_colonne, colonna_da_escludere)
job_senza_id <- job[, colonne_da_mantenere, drop = FALSE]
colnames(job)
 [1] "ID"         "Risk"       "Z"          "D"          "sex"        "age"
            "race"       "nonmarried" "alcohol6"
[10] "employ6"    "depress6"
colnames(job_senza_id)
 [1] "Risk"       "Z"          "D"          "sex"        "age"        "race
    "       "nonmarried" "alcohol6"   "employ6"
[10] "depress6"
```

4.2 NORMALIZZAZIONE DATI

Dopodichè, si è normalizzato i dati con cui si è dovuto lavorare in tipo *categorico*, questo perchè, in **R**, il metodo *hc()*, il quale crea una rete bayesiana a partire dai dati, non accetta valori di tipo *integer*.

```
# Normalizzazione dati
job_senza_id$Risk <- as.factor(job_senza_id$Risk)
job_senza_id$Z <- as.factor(job_senza_id$Z)
job_senza_id$sex <- as.factor(job_senza_id$sex)
job_senza_id$race <- as.factor(job_senza_id$race)
job_senza_id$nonmarried <- as.factor(job_senza_id$nonmarried)
job_senza_id$employ6 <- as.factor(job_senza_id$employ6)
job_senza_id$depress6 <- as.factor(job_senza_id$depress6)
```

4.3 CREAZIONE RETE BAYESIANA

Successivamente, si è creata la rete bayesiana a partire dai dati normalizzati presenti nel *dataset* ridotto, creato il plot e stampati a schermo gli archi presenti.

```
# Creazione rete bayesiana
job_bn <- hc(job_senza_id)
plot(job_bn)
```

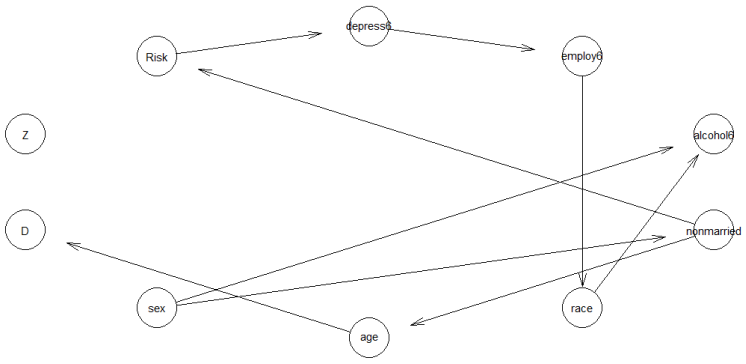

Figura 7: Rete bayesiana creata a partire dai dati presenti nel *dataset*.

```
print(arcs(job_bn))

     from         to
[1,] "Risk"       "depress6"
[2,] "nonmarried" "age"
[3,] "sex"        "alcohol6"
[4,] "depress6"   "employ6"
[5,] "sex"        "nonmarried"
[6,] "race"       "alcohol6"
[7,] "employ6"    "race"
[8,] "nonmarried" "Risk"
[9,] "age"        "D"
```

4.3.1 *Studio della rete Bayesiana*

Partendo dalla variabile Sex, è possibile che persone dello stesso sesso consumino maggiori dosi di Alcol e per questo vi è l'arco tra le due variabili Sex e Alcohol6; inoltre, è possibile che, nel paese del caso di studio, ci sia una maggioranza di un sesso rispetto all'altro, pertanto, è probabile che il sesso preponderante non sia sposato, e quindi la connessione anche con Nonmarried.
Risk è connesso a Depress6, questo perchè essere disoccupati ad alto rischi, può comportare un livello alto di depressione nel futuro (6 mesi).
Depress6 è connessa ad Employ6, poichè un alto livello di depressione può causare problemi con la società, mancanza di voglia ed energie per cercare lavoro e rendersi socialmente occupati.
Invece, Race è connessa ad Alcohol6 perchè, è possibile che ci siano alcuni gruppi di persone presenti all'interno della stessa etnia, i quali non ammettono l'uso di alcolici.
Age è connessa a D, poichè, per esempio, in età giovanile non si è abbastanza maturi da decidere di intraprendere un percorso di trattamento oppure in età adulta non si ha nè tempo, nè voglia di farlo.
Inoltre, Nonmarried è connessa a Risk, questo perchè, è possibile che dato l'essere non sposato possa comportare alcuni problemi, come quello dell'autostima e/o insicurezza.
Successivamente, Nonmarried è connesso anche ad Age, ma questo collegamento è improbabile, poichè implica che l'essere non sposati determini l'età di una persona.
Stesso discorso per Employ6, il quale è connesso a Race. Questo implica che l'etnia di una persona è determinata dall'essere reimpiegati dopo 6

mesi, il che è assurdo.
Pertanto, nella sezione successiva si mostra come eliminare le dipendenze per ottenere una nuova rete bayesiana più significativa.

4.4 AGGIORNAMENTO DELLA RETE BAYESIANA

Come spiegato, la rete bayesiana ottenuta risulta essere *naive*, questo perchè Nonmarried è connesso ad Age ed Employ6 a Race.
Successivamente, viene mostrato il codice utilizzato per ottenere una rete bayesiana più significativa e fatto un commento a riguardo.

```
# Matrice con tutti 0
blM <- matrix(0, nrow=10, ncol=10)
rownames(blM) <- colnames(blM) <- names(job_senza_id)
block <- c(2, 1, 1, 2, 2, 2, 2, 0, 0, 0)

# M[i,j] = 1 -> arco da i a j
for (b in 0:2) blM[block < b, block==b] <- 1

# Creazione Blacklist -> archi che non si desidera
blackL <- data.frame(get.edgelist(as(blM, "igraph")))
names(blackL) <- c("from", "to")

# Creazione rete bayesiana dai dati, specificando quali archi non si
    desidera
job_bn <- hc(job_senza_id, blacklist = blackL)
plot(job_bn)
```

Innanzitutto, si è creata la matrice 10x10, dove 10 indica il numero delle variabili totali, inizializzata a 0, in cui i nomi delle righe e delle colonne corrispondono ai nomi delle variabili presenti nel *dataset*.
Dopodichè, si è associata ad ogni variabile una categoria:

- 0: se la variabile corrisponde al post-trattamento.
- 1: se la variabile non corrisponde nè al pre, nè al post-trattamento.
- 2: se la variabile corrisponde al pre-trattamento.

Questo è servito per impostare a 1 l'elemento di riga i e colonna j, della matrice creata, il quale indica che l'elemento i è connesso all'elemento j ($i \rightarrow j$). Infatti, così facendo, questi archi sono stati aggiunti alla *Blacklist*, ovvero, alla lista degli archi da non considerare per la creazione della rete bayesiana.
Dato che si è ottenuta una rete simile a quella di partenza, ma con ancora

presente l'arco Nonmarried→Age, si è specificato l'arco in questione nella BL, per non considerarlo e creare la nuova rete bayesiana.

```
# Aggiungere alla BL l'arco: Nonmarried -> Age
blackL <- rbind(blackL, c("nonmarried", "age"))

# Nuova rete bayesiana da BL aggiornata
job_bn <- hc(job_senza_id, blacklist = blackL)
plot(job_bn)
```

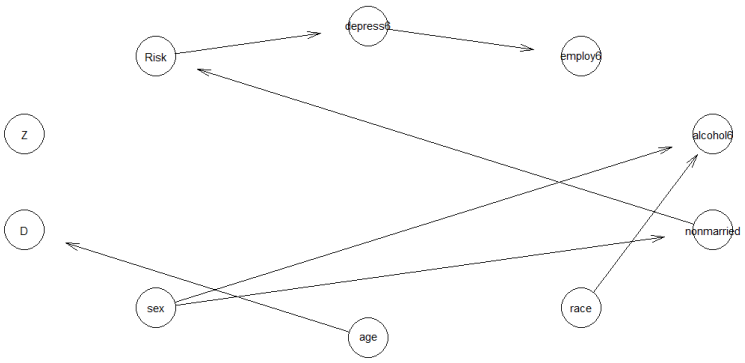

Figura 8: Rete bayesiana creata a partire dai dati presenti nel *dataset*, specificando quali archi non considerare presenti nella *Blacklist*.

4.5 REGRESSIONI BASATE SULLA RETE BAYESIANA

Proseguendo, si sono fatte alcune regressioni logistiche basate sul DAG ottenuto dalla rete bayesiana precedentemente creata.

```
# Regressioni basate sulla rete bayesiana
job_bn_employ6 <- glm(employ6 ~ depress6, family = binomial, data = job_
    senza_id)
summary(job_bn_employ6)

Call:
glm(formula = employ6 ~ depress6, family = binomial, data = job_senza_id)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.26488    0.08511   3.112  0.00186 **
```

```
depress61    0.51825    0.12924   4.010 6.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1396.4  on 1052  degrees of freedom
Residual deviance: 1380.1  on 1051  degrees of freedom
AIC: 1384.1

Number of Fisher Scoring iterations: 4

print(BIC(job_bn_employ6))
[1] 1394.035
```

Si osserva che, l'unico vertice connesso ad Employ6, cioè Depress6, risulta essere fortemente significativo. Inoltre, il fatto che sia positivo, indica che influenza positivamente la variabile dipendente.

```
job_bn_depress6 <- glm(depress6 ~ Risk, family = binomial, data = job_senza
    _id)
summary(job_bn_depress6)

Call:
glm(formula = depress6 ~ Risk, family = binomial, data = job_senza_id)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.35516    0.08012   4.433  9.3e-06 ***
Risk1       -1.32151    0.13651  -9.681  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1455.0  on 1052  degrees of freedom
Residual deviance: 1354.2  on 1051  degrees of freedom
AIC: 1358.2

Number of Fisher Scoring iterations: 4

print(BIC(job_bn_depress6))
[1] 1368.126
```

Anche in questo caso, l'unico vertice connesso a Depress6, cioè Risk, risulta essere fortemente significativo per la depressione, ma al contrario del caso precedente, Risk influenza negativamente questa patologia.

Successivamente, viene mostrata la dipendenza di Alcohol6 da Sex e Race.

```
job_bn_alcohol6 <- glm(alcohol6 ~ sex + race, family = binomial, data = job
    _senza_id)
summary(job_bn_alcohol6)

Call:
glm(formula = alcohol6 ~ sex + race, family = binomial, data = job_senza_id
    )

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.52067    0.09802  -5.312 1.09e-07 ***
sex1         0.69790    0.12726   5.484 4.16e-08 ***
race1        0.87829    0.16902   5.196 2.03e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1459.7  on 1052  degrees of freedom
Residual deviance: 1399.3  on 1050  degrees of freedom
AIC: 1405.3

Number of Fisher Scoring iterations: 4

print(BIC(job_bn_alcohol6))
[1] 1420.175
```

Difatti, com'è possibile notare, Sex e Race sono fortemente significative per l'intercetta.
I coefficienti positivi di Sex1 e Race1 indicano che queste variabili influenzano positivamente il livello di alcol passati i 6 mesi.

Fino ad ora le regressioni si sono basate su tutte e sole le dipendenze fornite dalla rete bayesiana. Adesso vengono mostrate alcune regressioni contenenti variabili che non sono state rappresentate nella rete, le quali, nonostante ciò, hanno più o meno importanza.

```
job_bn_employ6 <- glm(employ6 ~ depress6 + age + race + Z, family =
    binomial, data = job_senza_id)
summary(job_bn_employ6)

Call:
glm(formula = employ6 ~ depress6 + age + race + Z, family = binomial,
    data = job_senza_id)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.594711   0.271726   5.869 4.39e-09 ***
depress61    0.558049   0.132239   4.220 2.44e-05 ***
age         -0.031881   0.006411  -4.972 6.61e-07 ***
race1       -0.663299   0.163523  -4.056 4.99e-05 ***
Z1          -0.052370   0.140311  -0.373    0.709
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1396.4  on 1052  degrees of freedom
Residual deviance: 1341.8  on 1048  degrees of freedom
AIC: 1351.8

Number of Fisher Scoring iterations: 4

print(BIC(job_bn_employ6))
[1] 1376.56
```

In questo caso, a differenza del precedente, si sono aggiunte al modello tre variabili non connesse con l'intercetta, per provare a verificare il loro impatto su quest'ultima.
Si nota che l'età ha un effetto marcato, fortemente significativo, al contrario del trattamento, il quale è per nulla significativo.
Il coefficiente positivo di Depress61 indica che la depressione influisce positivamente sull'essere reimpiegati o meno, mentre per l'età e per l'etnia non è così, anzi: come si può osservare, esse hanno un coefficiente negativo, pertanto influenzano negativamente la variabile dipendente.

```
job_bn_depress6 <- glm(depress6 ~ Risk + employ6 + D, family = binomial,
    data = job_senza_id)
summary(job_bn_depress6)

Call:
glm(formula = depress6 ~ Risk + employ6 + D, family = binomial,
    data = job_senza_id)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.005049   0.116603   0.043   0.9655
Risk1       -1.335575   0.138253  -9.660  < 2e-16 ***
employ61     0.544729   0.136313   3.996 6.44e-05 ***
D            0.196596   0.078486   2.505   0.0122 *
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1455  on 1052  degrees of freedom
Residual deviance: 1332  on 1049  degrees of freedom
AIC: 1340

Number of Fisher Scoring iterations: 4

print(BIC(job_bn_depress6))
[1] 1359.848
```

Anche in quest'ultimo caso si sono aggiunte variabili non direttamente connesse all'intercetta, come D e Employ6, dove quest'ultima è fortemente significativa, mentre la prima quasi per niente.
Il coefficiente positivo di Employ61 pone l'attenzione sul fatto che influenza positivamente la depressione; cosa opposta per la variabile Risk1, dato che il suo coefficiente è negativo.

4.6 INTERROGARE LA RETE BAYESIANA

In questa sezione si affronta la parte di interrogazione della rete bayesiana, ma per questo si è dovuto creare una nuova rete affinchè si potesse utilizzare il metodo apposito.
Per fare questo, si è partiti dalla creazione del DAG, per poi creare la nuova rete, per, infine, interrogarla.

4.6.1 *Creazione del DAG*

Dato che la rete bayesiana creata è un oggetto di tipo *bn*, si è creato un DAG basato sulla sua struttura, in modo tale da creare una nuova rete di tipo *grain*.

```
# Creazione DAG
job_dag <- dag(~ D * age,
               ~ Risk * nonmarried,
               ~ depress6 * Risk,
               ~ employ6 * depress6,
               ~ alcohol6 * race,
               ~ alcohol6 * sex,
               ~ nonmarried * sex
               )
```

4.6.2 *Creazione della nuova Rete a partire dal DAG*

La nuova rete si è costruita basandosi sul DAG appena creato e utilizzando i dati presenti nel *dataset* in cui manca la colonna ID.

```
# Rete bayesiana dal DAG
job_bn_da_dag <- grain(job_dag, data = job_senza_id)
job_bn_da_dag <- compile(job_bn_da_dag, propagate = TRUE, smooth=0.1)
```

4.6.3 *Probabilità marginali*

Le prime interrogazioni fatte sono state quelle riferite alle probabilità marginali, ovvero quelle basate sul sapere quale sia la probabilità che una variabile possa assumere determinati valori, in maniera indipendente dalle altre.

```
# Probabilita' marginali
querygrain(job_bn_da_dag, nodes = c("sex"), type = "marginal")
        0         1
0.4634378 0.5365622

querygrain(job_bn_da_dag, nodes = c("Risk"), type = "marginal")
        0         1
0.6106363 0.3893637
```

4.6.4 *Probabilità congiunte*

In seconda battuta, si è interrogata la rete in merito alle probabilità congiunte, ovvero le probabilità che due eventi accadano contemporaneamente.

```
# Probabilita' congiunte
querygrain(job_bn_da_dag, nodes = c("sex", "nonmarried"), type = "joint")

   nonmarried
sex         0         1
  0 0.2402659 0.2231719
  1 0.2146249 0.3219373

querygrain(job_bn_da_dag, nodes = c("race", "alcohol6"), type = "joint")

    alcohol6
race          0         1
```

```
   0 0.43475215 0.3772137
   1 0.06243191 0.1256023

querygrain(job_bn_da_dag, nodes = c("Risk", "depress6"), type = "joint")
    depress6
Risk         0         1
   0 0.2516619 0.3589744
   1 0.2820513 0.1073124
```

Dal primo caso si può dedurre che non ci siano differenze marcate tra il genere e l'essere sposato o meno. Tuttavia, sembrerebbe che sia più marcato l'essere femmina e non sposata.
Invece, per il secondo, risulta evidente che la probabilità di essere un soggetto di etnia bianca, con alti livelli di alcol, è più marcata del caso di etnia diversa.
Infine, per l'ultimo caso, si nota che avendo un rischio basso, si ha anche un livello di depressione inferiore, al contrario di coloro che hanno un rischio alto e un più elevato livello di depressione.

4.6.5 *Probabilità condizionali*

Per ultimo, il caso delle probabilità condizionate, ovvero quelle probabilità in cui il verificarsi di un evento y influenza un risultato x ($P(x \mid y)$).

```
# Probabilita' condizionali
querygrain(job_bn_da_dag, nodes = c("employ6", "depress6"),
type = "conditional")

       depress6
employ6         0         1
      0 0.4341637 0.3136456
      1 0.5658363 0.6863544

querygrain(job_bn_da_dag, nodes = c("depress6", "Risk"),
type = "conditional")

        Risk
depress6         0         1
       0 0.4121306 0.7243902
       1 0.5878694 0.2756098
```

Dal primo risultato si evince che la probabilità di essere reimpiegati, dato un basso livello di depressione, è molto più alta rispetto a coloro che non

sono reimpiegati, ma con lo stesso livello patologico. Invece, non vi è molta differenza per coloro che hanno un alto livello di depressione. Infine, la probabilità di avere un alto livello di depressione essendo un soggetto disoccupato ad alto rischio, è elevata. Mentre, la probabilità di avere un livello di depressione basso, è poco influenzato dal fatto di essere un soggetto ad alto rischio.

5

CONCLUSIONI

In conclusione, l'analisi preliminare ha indicato che il trattamento potrebbe non avere rilevanza significativa nei dati esaminati. Questa osservazione è stata ulteriormente supportata dall'analisi dei modelli logistici (*Interpretazione coefficienti per il reimpiego* e *Interpretazione coefficienti per la depressione*) e delle reti bayesiane (*Reimpiego* e *Depressione*), i quali non includono il trattamento come fattore significativo. Di conseguenza, è possibile concludere che, sulla base delle evidenze statistiche raccolte, il trattamento non appare essere un elemento determinante o influente nelle relazioni studiate. Questi risultati pongono l'accento sull'importanza di un'approfondita analisi statistica per comprendere la reale incidenza di variabili specifiche e forniscono indicazioni cruciali per decisioni future o ulteriori ricerche nell'ambito di questo studio.