Tarea 4

Mario Becerra 124362

09/02/2015

En este ejercicio utilizaremos la base de datos *insurance* incluida en el paquete *bnlearn*.

De acuerdo a un experto se ajustó una red con las siguientes consideraciones:

- Edad debe ser una variable raíz (no puede haber aristas con dirección a edad).
- La única determinante de nivel socioeconómico puede ser la edad, en caso de que sea necesario.
- El costo médico es independiente de la edad y nivel socieconómico dado las características del accidente y el tipo de coche (en particular, no puede haber aristas entre edad/nivel socioeconómico y costo médico).
- No puede haber aristas de existencia de bolsa de aire hacia calidad de conductor, historia de manejo o accidente, ni tampoco de accidente hacia tipo de coche o hacia año del coche.

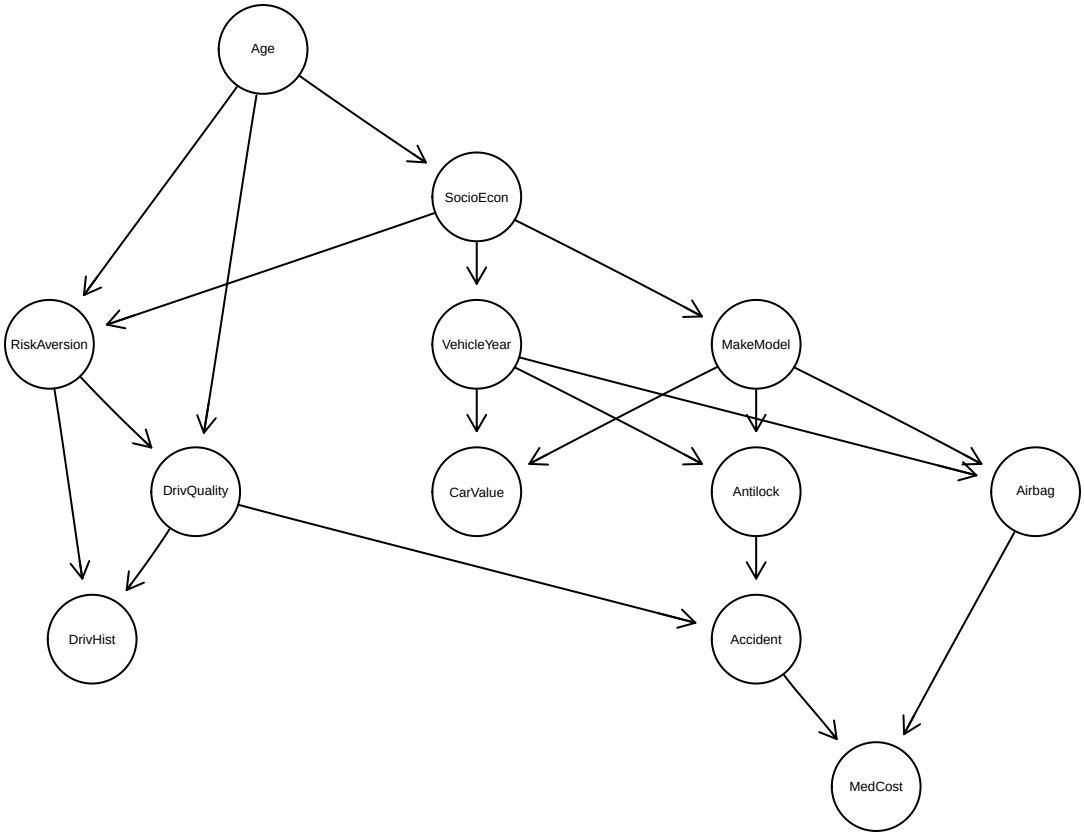
La red es la siguiente:

```
library(bnlearn)
library(dplyr)
```

```
set.seed(128)

vars <- c('Age', 'RiskAversion', 'VehicleYear', 'Accident', 'MakeModel',
          'DrivQuality', 'Airbag', 'DrivHist', 'SocioEcon', 'Antilock',
          'MedCost', 'CarValue')
blacklist_1 <- data.frame(from = vars[-1], to = 'Age')
blacklist_2 <- data.frame(from = c('MedCost', 'MedCost', 'Accident', 'Accident'),
                          to = c('Age', 'SocioEcon', 'MakeModel', 'DrivQuality'))
blacklist_3 <- data.frame(from = c('Age', 'SocioEcon'), to = c('MedCost'))
blacklist_4 <- data.frame(from = vars[-1], to = 'SocioEcon')
blist <- rbind(blacklist_1, blacklist_2, blacklist_3, blacklist_4)
net_insurance <- hc(insurance[, vars], score = 'bic', blacklist = blist)
graphviz.plot(net_insurance)
```

```
## Loading required namespace: Rgraphviz
```

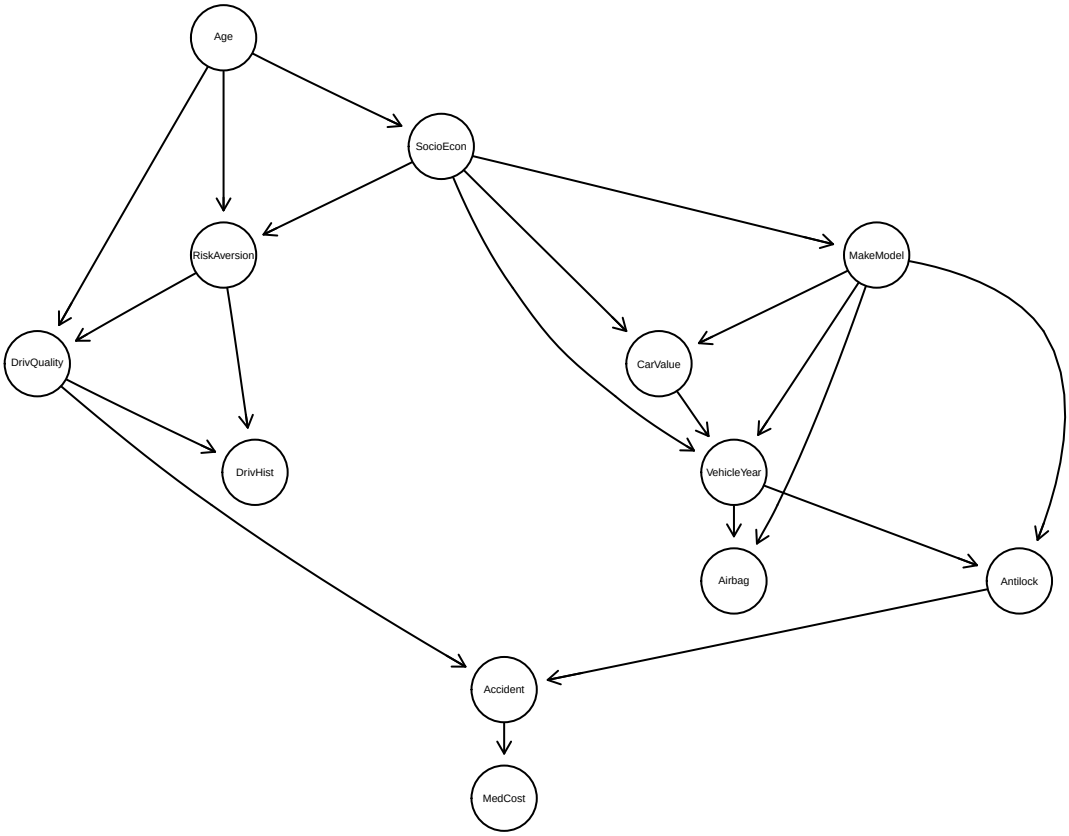


1. Usa máxima verosimilitud para estimar los parámetros de los modelos locales, utiliza la base de datos *insurance_sub* que se crea con las siguientes líneas:

```
set.seed(3656723)
insurance_sub <- sample_n(insurance, 3000)
```

Primero creamos la red correspondiente a *insurance_sub*.

```
net_insurance_sub <- hc(insurance_sub[, vars],  score = 'aic', blacklist = blist)
graphviz.plot(net_insurance_sub)
```

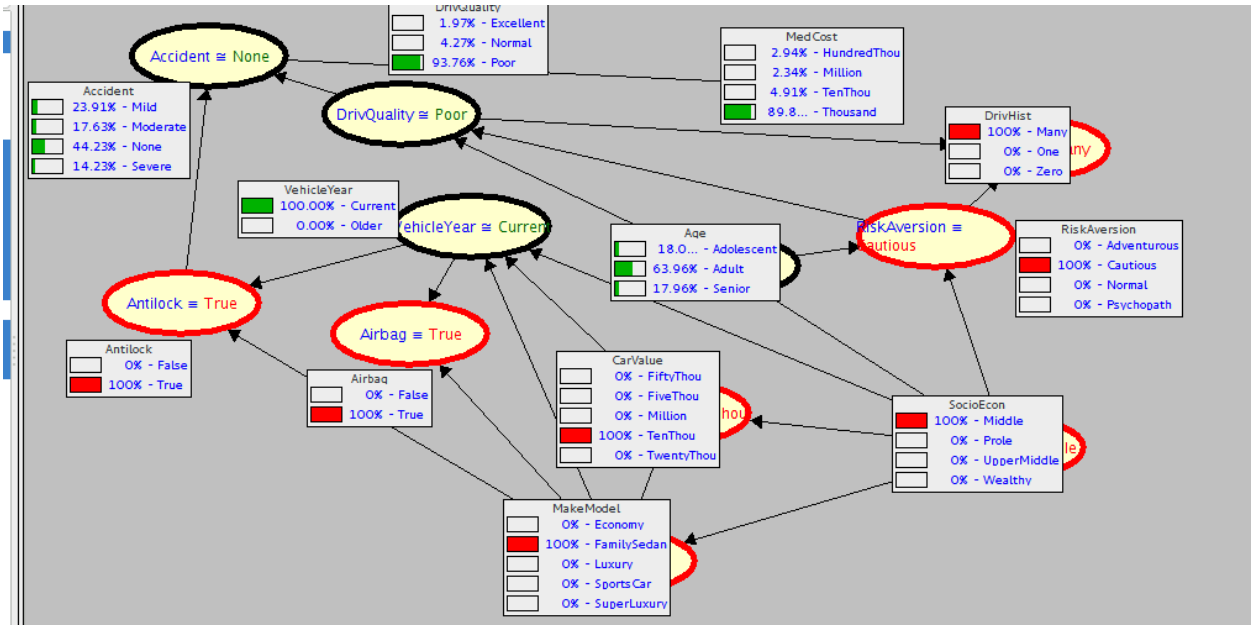


Calculando por máxima verosimilitud:

```
fit_insurance_mle <- bn.fit(net_insurance_sub, data = insurance_sub[,vars], method = 'mle')
write.net("./insurance_sub.net", fit_insurance_mle)
```

2. Exporta la red a SAMIAM y realiza algunos querys, ¿es necesario modificar alguna tabla de probabilidad condicional? es decir, ¿las estimaciones son ruidosas para algún nodo?

Sí, es necesario modificar por lo menos una probabilidad condicional, pues, como se puede ver en la figura, bajo las condiciones seleccionadas, se le asigna probabilidad cero a un coche viejo.



3. Utiliza un modelo logístico para modelar el nodo Antilock y una logística multinomial (por ejemplo de glmnet) para describir la relación VehicleYear -> CarValue <- MakeModel. Explica qué relaciones de independencia y dependencia describe este colisionador. Por ejemplo, si conozco que el valor del coche es de TenThou, ¿qué información del modelo de coche (MakeModel) aporta saber que el coche es nuevo Current?

```
mod1 <- glm(Antilock ~ VehicleYear + MakeModel, family='binomial', data=insurance_sub)
summary(mod1)
```

```
##
## Call:
## glm(formula = Antilock ~ VehicleYear + MakeModel, family = "binomial",
##     data = insurance_sub)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.89921  -0.09352  -0.00013  -0.00001   2.26875
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -18.5740   431.4217  -0.043    0.966
## VehicleYearOlder       -4.9578     0.2826 -17.545   <2e-16 ***
## MakeModelFamilySedan   18.1018   431.4217   0.042    0.967
## MakeModelLuxury        22.7616   431.4219   0.053    0.958
## MakeModelSportsCar     21.0375   431.4218   0.049    0.961
```

```
## MakeModelSuperLuxury   39.1401  7251.2386   0.005    0.996
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2968.84  on 2999  degrees of freedom
## Residual deviance:  923.68  on 2994  degrees of freedom
## AIC: 935.68
##
## Number of Fisher Scoring iterations: 19
```

Al parecer la variable *MakeModel* no es significativa. Probamos un nuevo modelo con la variable *VehicleYear* únicamente.

```
mod2 <- glm(Antilock ~ VehicleYear, family='binomial', data=insurance_sub)
summary(mod2)
```

```
##
## Call:
## glm(formula = Antilock ~ VehicleYear, family = "binomial", data = insurance_sub)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2174  -0.2034  -0.2034  -0.2034   2.7886
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.09367    0.06188   1.514     0.13
## VehicleYearOlder -3.96122    0.17132 -23.122   <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: 2968.8  on 2999  degrees of freedom
## Residual deviance: 1839.4  on 2998  degrees of freedom
## AIC: 1843.4
##
## Number of Fisher Scoring iterations: 6
```

Según este modelo, hay una relación negativa entre si el vehículo es viejo si el auto tiene *antilock*.

Ajustamos ahora un modelo para *CarValue*.

```
mod3 <- glm(CarValue ~ VehicleYear + MakeModel, family='binomial', data=insurance_sub)
summary(mod3)
```

```
##
## Call:
## glm(formula = CarValue ~ VehicleYear + MakeModel, family = "binomial",
##     data = insurance_sub)
##
```

```
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.319   0.000   0.000   0.000   1.153
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              20.7770  1255.7265   0.017    0.987
## VehicleYearOlder          3.7154     0.4404   8.437   <2e-16 ***
## MakeModelFamilySedan      0.5838  1777.8307   0.000    1.000
## MakeModelLuxury         -24.4337  1255.7265  -0.019    0.984
## MakeModelSportsCar      -18.9900  1255.7265  -0.015    0.988
## MakeModelSuperLuxury      1.7891 19716.0234   0.000    1.000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1901.71  on 2999  degrees of freedom
## Residual deviance:  281.25  on 2994  degrees of freedom
## AIC: 293.25
##
## Number of Fisher Scoring iterations: 21
```

En este modelo tampoco parece haber relaciones significativas.