CAT - 05) Modelos loglineales

Los datos de preferencias de automóviles "Car preferences.xls", tienen tres variables de clasificación: sex, age y response que se refiere a que tan importante: no/little, important, very important es que el auto tenga ayuda o asistencia de manejo, tal como sensores/cámaras para estacionarse, frenar al acercarse mucho a otro auto.

```
# Load database
library(readxl)
df_data = read.csv("car_preferences.csv")

# Transform columns into factors, define levels order
sex_levels = c("women", "men")
age_levels <- c("18-23", "24-40", "> 40")
response_levels <- c("no/little", "important", "very important")

# Asign factors to dataframe
df_data$sex = factor(df_data$sex, sex_levels)
df_data$age <- factor(df_data$age, levels=age_levels)
df_data$response <- factor(df_data$response, levels=response_levels)

# Display database
df_data
```

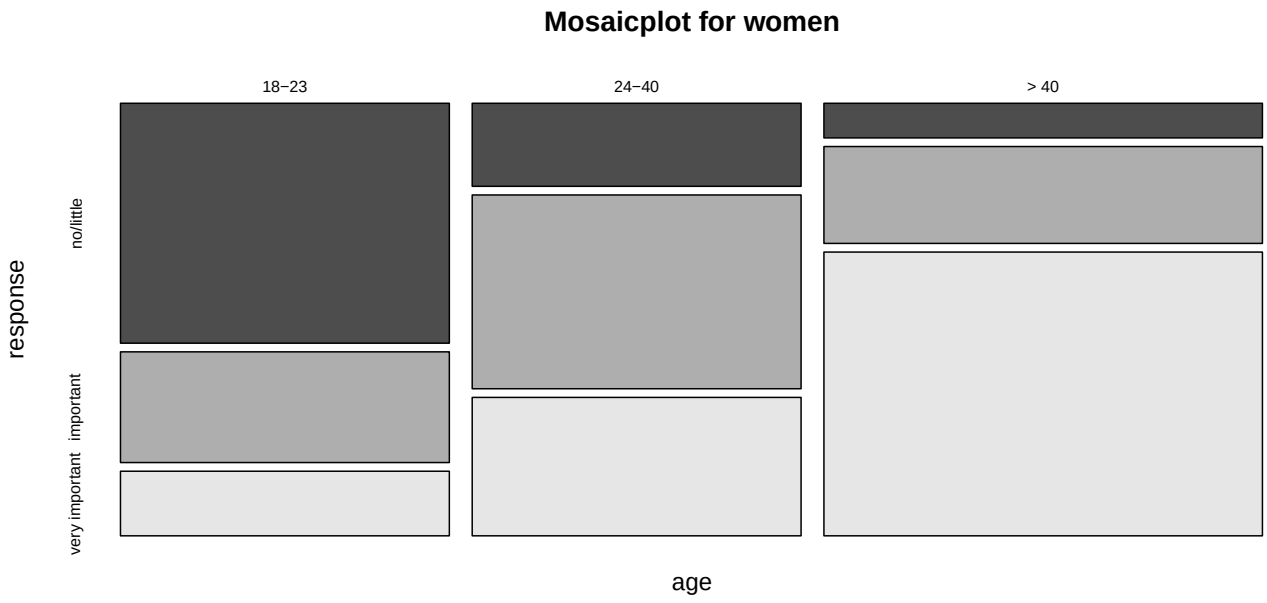
```
##      sex   age       response frequency
## 1  women 18-23      no/little        26
## 2  women 18-23      important        12
## 3  women 18-23 very important         7
## 4  women 24-40      no/little         9
## 5  women 24-40      important        21
## 6  women 24-40 very important        15
## 7  women  > 40      no/little         5
## 8  women  > 40      important        14
## 9  women  > 40 very important        41
## 10   men 18-23      no/little        40
## 11   men 18-23      important        17
## 12   men 18-23 very important         8
## 13   men 24-40      no/little        17
## 14   men 24-40      important        15
## 15   men 24-40 very important        12
## 16   men  > 40      no/little         8
## 17   men  > 40      important        15
## 18   men  > 40 very important        18
```

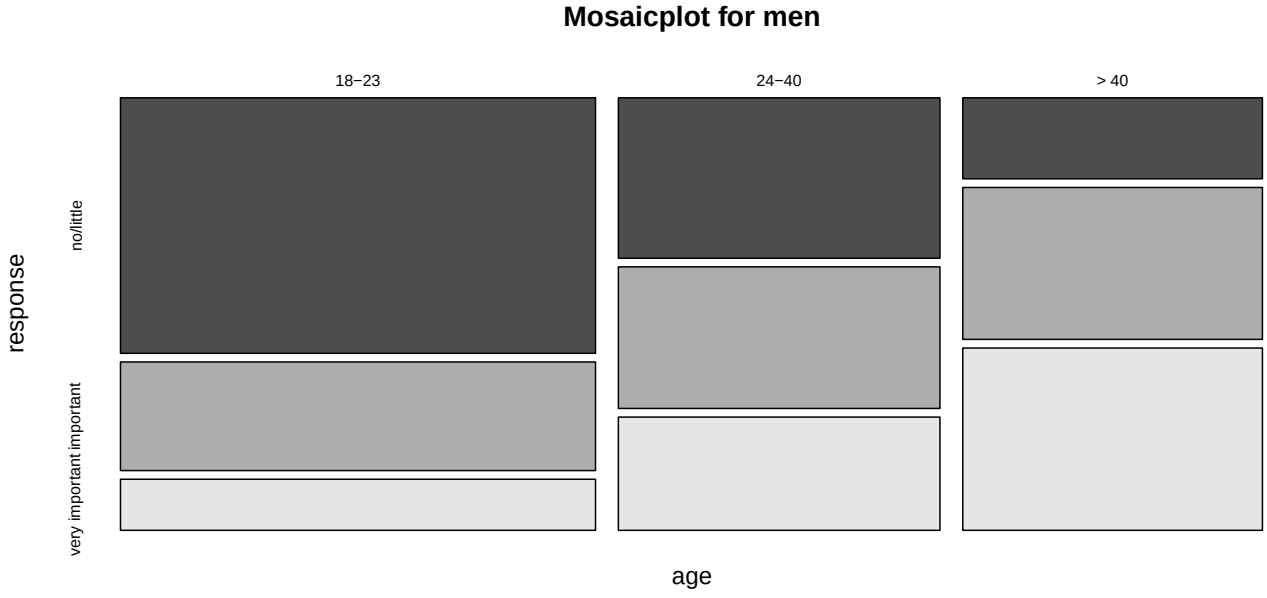
Conteste lo siguiente:

9.a) Haz el mosaic plot de los datos.

```
# Unique values for sex factors
sexes = unique(df_data$sex)

# Create a 2x1 panel to print mosaicplots
par(mfrow=c(2, 1))

# Iterate for both sexes
for (sex in sexes) {
  # Slice df for current sex
```

```
  df_temporal = df_data[df_data$sex==sex,]

  # Transform sliced df into a contingency table
  cont_table = xtabs(frequency ~ age+response, data=df_temporal)

  # Plot mosaic plot for current sex of age vs reposne
  mosaicplot(cont_table, main=paste("Mosaicplot for", sex), color=TRUE)
}
```





9.b) Para qué grupo de edad es muy importante tener asistencia.

En los mosaicplot se observa que tanto para hombres como para mujeres, mientras más edad tengan, es más importante tener asistencia, pues existe una progresión con la edad en la proporción de frecuencias. En ambos sexos, para los jóvenes el grupo minoritario es el que lo considera importante, pero para los mayores de 40 éste grupo es el mayoritario.

9.c) Ajusta los tres modelos loglineales de una variable explicativa, tres con una interacción, tres con dos interacciones, modelo de asociación homogénea y el saturado.

Para este ejercicio las variables son codificadas en:

s: sex ; a: age ; r: response

```
# Single variable models
model_s = glm(frequency~sex, family=poisson, data=df_data)
model_a = glm(frequency~age, family=poisson, data=df_data)
model_r = glm(frequency~response, family=poisson, data=df_data)

# Mutual independence - p_ijk = p_i * p_j * p_k
# All variables are independent of one another
model_s_a_r = glm(frequency~sex + age + response,
                  family=poisson, data=df_data)

# Joint independence - p_ijk = p_ij * p_k
# There is an interaction between A and B that is unaffected by C.
# C is independent of A and B.
model_sa_r = glm(frequency~sex + age + response + sex*age,
                 family=poisson,data=df_data)
model_sr_a = glm(frequency~sex + age + response + sex*response,
                 family=poisson,data=df_data)
model_ar_s = glm(frequency~sex + age + response + age*response,
                 family=poisson,data=df_data)

# Conditional independence - p_ijk = p_ij * p_ik / p_i
# Conditional independence between B and C, controlling for A.
# Within each level of A, the other variables B and C are independent
model_sa_sr = glm(frequency~sex + age + response + sex*age + sex*response,
                  family=poisson,data=df_data)
model_sa_ar = glm(frequency~sex + age + response + age*sex + age*response,
                  family=poisson,data=df_data)
model_sr_ar = glm(frequency~sex + age + response + response*sex + response*age,
                  family=poisson,data=df_data)

# Homogeneous association
# There are intereactions between each pair of variables, but each interaction
# is unnafected by the category of the third variable
model_homog = glm(frequency~sex + age + response + sex*age + sex*response +
                      age*response, family=poisson,data=df_data)

# Saturated model
# The interaction between any two variables is affected by the third variable
model_sat = glm(frequency~sex + age + response + sex*age + sex*response +
                age*response + sex*age*response, family=poisson,data=df_data)
```

9.d) Haz una tabla con las devianzas, grados de libertad, AIC, términos incluidos en cada modelo ajustado y tipo de independencia asociada a cada modelo.

En la tabla debajo se muestran los resultados obtenidos al ajustar los distintos modelos.

Clave	Terminos incluidos	Tipo de independencia	Devianza (nula = 91.903)	Grados de libertad	AIC
s	sex (s)	Indep. marginal	91.903	16	177.13
a	age (a)	Indep. marginal	89.667	15	176.9
r	response (r)	Indep. marginal	91.279	15	178.51
s_a_r	s+a+r	Indep. mutua	89.044	12	182.27
sa_r	s+a+r + sa	Indep. conjunta	81.781	10	179.01
sr_a	s+a+r + sr	Indep. conjunta	76.781	10	174.01
ar_s	s+a+r + ar	Indep. conjunta	17.707	8	118.94
sa_sr	s+a+r + sa+sr	Indep. condicional	69.518	8	170.75
sa_ar	s+a+r + sa+ar	Indep. condicional	10.444	6	115.67
sr_ar	s+a+r + sr+ar	Indep. condicional	5.4446	6	110.67
homog	s+a+r + sa+sr+ar	Asoc. homogénea	3.9387	4	113.17
sat	s+a+r + sa+sr+ar + sar	Modelo saturado	0	0	117.23

9,e) ¿Qué modelo crees que se ajusta mejor? Justifica tu respuesta.

El método más sencillo para selección de modelo, conociendo la información presentada en la tabla anterior, es utilizar el Criterio de Información de Akaike (AIC). El modelo con el AIC más bajo es el que mejor explica los datos, castigando por el número de parámetros estimados. El criterio puede interpretarse como una "distancia", el modelo con la menor distancia es el óptimo.

Por lo tanto, el modelo que se cree que ajusta mejor es el sr_ar (AIC=110.67), que es de la forma:

$$ln(\mu_{ijk}) = \lambda + \lambda_i^{sex} + \lambda_j^{age} + \lambda_k^{response} + \lambda_{ij}^{sex-response} + \lambda_{ik}^{age-response}$$

Este modelo considera dos interacciones, en ambas la variable *response* está presente. Esto implica que *sex* y *age* son condicionalmente independientes, condicionado por *response*.

9.f) Comenta sobre los parámetros estimados del modelo elegido.

En el script debajo se muestran los parámetros obtenidos.

Se puede observar que todas las interacciones de *age* con *response* son significativas, por lo que confirmamos la hipótesis que se tuvo al principio, que entre más edad, más importancia se le toma a la asistencia de manejo. Las categorías de referencia son *18-23* y *no/little*.

En el caso de la interacción (*age> 40:responsevery important*), el parámetro estimado es 2.9942. El cuál es mayor al de la misma interacción con *age: 24-40*. Entonces los momios son $e^{2.9942} = 19.96$, lo que indica que la celda tiene una frecuencia más grande de la que se esperaría si *age* y *response* fueran independientes.

En otro cas, la interacción *sexmen:responsevery important* tiene un parámetro estimado negativo, lo que implica que se reduce la frecuencia esperada si *sex* y *response* fueran independientes. Debido a que la variable de referencia es *women*, esto quiere decir que para los hombres es más importante tener asistencias de manejo.

```
summary(model_sr_ar)
```

```
##
## Call:
## glm(formula = frequency ~ sex + age + response + response * sex +
##     response * age, family = poisson, data = df_data)
##
```

```
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       3.2246     0.1750  18.425  < 2e-16 ***
## sexmen                            0.4855     0.2010   2.416 0.015694 *
## age24-40                         -0.9316     0.2315  -4.023 5.74e-05 ***
## age> 40                          -1.6247     0.3034  -5.354 8.59e-08 ***
## responseimportant                -0.5504     0.2752  -2.000 0.045510 *
## responsevery important           -0.9885     0.3214  -3.076 0.002097 **
## sexmen:responseimportant         -0.4855     0.2880  -1.686 0.091824 .
## sexmen:responsevery important    -0.9911     0.2874  -3.449 0.000563 ***
## age24-40:responseimportant        1.1478     0.3404   3.372 0.000747 ***
## age> 40:responseimportant         1.6247     0.4013   4.049 5.15e-05 ***
## age24-40:responsevery important   1.5193     0.3966   3.831 0.000128 ***
## age> 40:responsevery important    2.9942     0.4192   7.143 9.10e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 91.9026  on 17  degrees of freedom
## Residual deviance:  5.4446  on  6  degrees of freedom
## AIC: 110.67
##
## Number of Fisher Scoring iterations: 4
```