Insurance

Carichiamo il dataset `insurance` relativo all'ammontare delle spese mediche di individui americani attualmente iscritti ad un piano assicurativo.

```
data = read.csv("insurance.csv")
```

Vediamo quali sono le variabili incluse nel dataset.

```
str(data)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr  "female" "male" "male" "male" ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

Si nota innanzitutto che il dataset è relativo a 1338 individui e contiene 7 variabili:

- `age`: età del beneficiario principale (variabile quantitativa);
- `sex`: sesso del contribuente assicurativo (fattore a due livelli, `female` e `male`);
- `bmi`: *Body Mass Index*, indice di massa corporea (variabile quantitativa);
- `children`: numero di figli coperti dall'assicurazione;
- `smoker`: se i singoli individui sono o meno fumatori (fattore a due livelli, `yes` e `no`);
- `region`: regione di provenienza del beneficiario negli Stati Uniti (fattore a 4 livelli, `southwest`, `southeast`, `northwest` e `northeast`);
- `charges`: spese mediche individuali addebitate alle compagnie assicurative relative ad un anno solare.

Dalla struttura di `data` osserviamo che le variabili categoriali non sono fattori. Procediamo dunque alla trasformazione di queste variabili.

```
library(dplyr)
data = data %>%
  mutate_if(is.character,as.factor)
str(data)
```

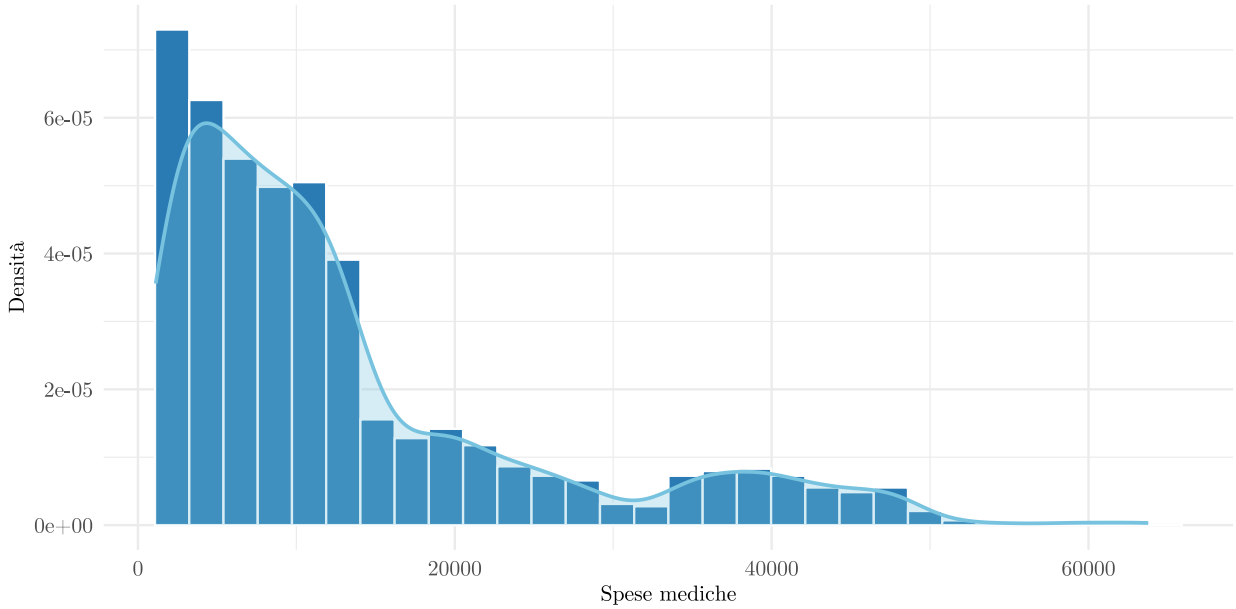
```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

Analisi esplorativa

Per comprendere meglio questo insieme di dati effettuiamo un'analisi esplorativa.

La variabile di interesse è `charges`, l'ammontare delle spese mediche degli individui addebitate alla compagnia assicurativa. Rappresentiamo la distribuzione di questa variabile.

```
library(ggplot2)
library(paletteer)
ggplot(data,aes(x = charges)) +
  geom_histogram(aes(y = ..density..),bins = 30,
                 fill = paletteer_c("ggthemes::Classic Blue",6)[4],
                 col = "white") +
  geom_density(col = paletteer_c("ggthemes::Classic Blue",6)[2],
               fill = paletteer_c("ggthemes::Classic Blue",6)[2],
               alpha = 0.3,size = 0.8) +
  labs(x = "Spese mediche",y = "Densità") +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 10))
```



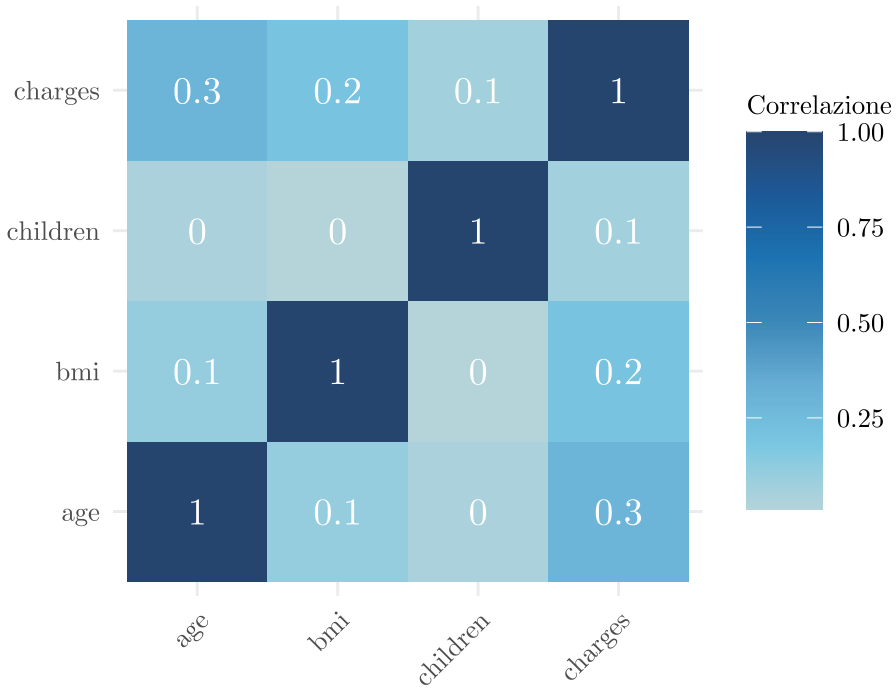
La distribuzione delle spese mediche è asimmetrica a destra e presenta una coda lunga. Si evidenziano inoltre diversi massimi locali.

Si può vedere che la maggior parte degli individui presenta spese mediche inferiori a 10.000 dollari. Tuttavia, ci sono alcuni individui che hanno spese molto elevate, fino a 60.000 dollari.

Valutiamo ora la correlazione tra le variabili. Considerato che non tutte le variabili sono quantitative consideriamo quindi solamente `age`, `bmi`, `children`, `charges`.

```
library(reshape2)
data_numeric = data[,sapply(data,is.numeric)]
corr = cor(data_numeric)
cor_melted = melt(corr)
```

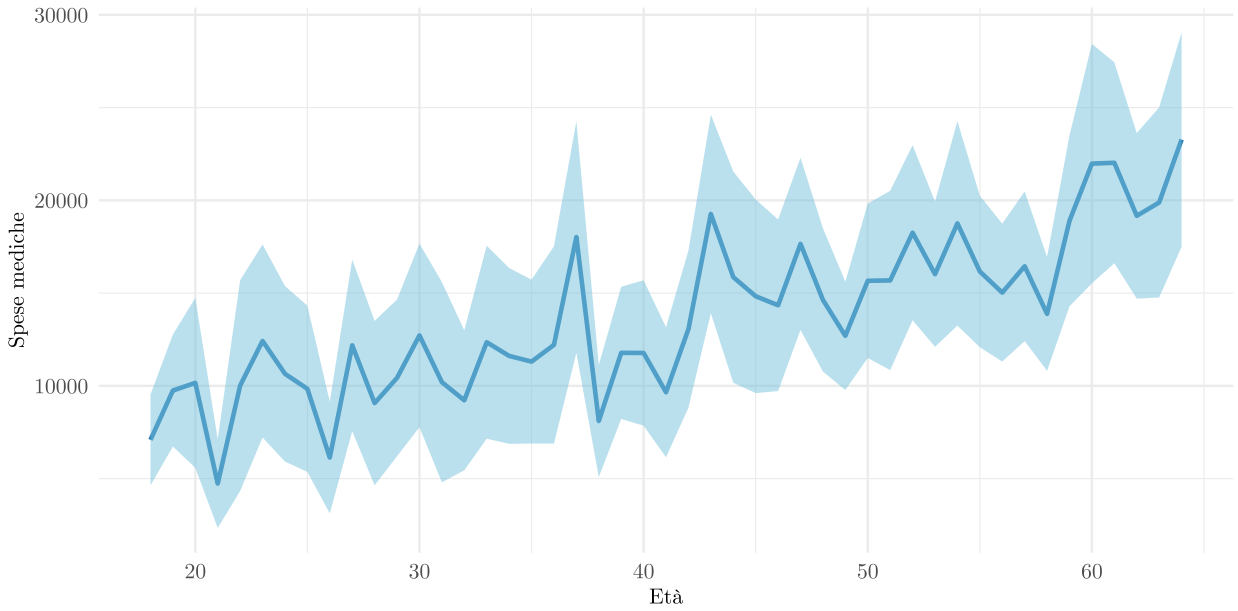
```r
cor_melted %>%
  ggplot(aes(Var1,Var2,fill = value)) +
  geom_tile() +
  scale_fill_paletteer_c(`"ggthemes::Classic Blue"`) +
  geom_text(aes(label = round(value,1)),family = "CMUSerif",size = 5,col = "white") +
  theme_minimal() +
  labs(fill = "Correlazione") +
  xlab("") +
  ylab("") +
  coord_fixed()+
  theme(axis.text.x = element_text(angle = 45,hjust = 1),
        text = element_text(family = "CMUSerif"),
        axis.title = element_text(size = 10),
        axis.text = element_text(size = 10),
        legend.text = element_text(size = 10),
        legend.title = element_text(size = 10),
        legend.key.size = unit(1,"cm"))
```



La nostra variabile di interesse, `charges`, risulta essere positivamente correlata con tutte le altre variabili quantitative. La variabile più legata alle spese mediche risulta l'età con una correlazione di 0.3.

Il valore positivo della correlazione tra `charges` e `age` ci informa sul fatto che all'aumentare dell'età anche l'ammontare delle spese mediche subirà una crescita. Visualizziamo graficamente questa tendenza.

```r
ggplot(data,aes(x = age,y = charges)) +
  stat_summary(fun = mean,geom = "line",
               col = paletteer_c("ggthemes::Classic Blue",6)[4],size = 1) +
  stat_summary(fun.data = mean_cl_normal,geom = "ribbon",
               fill = paletteer_c("ggthemes::Classic Blue",6)[2],alpha = 0.5) +
  labs(x = "Età",y = "Spese mediche") +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 10))
```



Nel grafico, la linea continua blu è relativa alla media della variabile `charges` mentre l'intervallo azzurro indica i valori minimi e massimi assunti da questa variabile per ogni valore di `age`.

Per le restanti variabili categoriali possiamo calcolare l'**indice di associazione** η^2, una misura che quantifica la forza dell'associazione di una variabile indipendente qualitativa sulla variabilità di una variabile dipendente quantitativa in contesti di analisi della varianza (ANOVA). In termini più semplici, η^2 rappresenta la proporzione della varianza totale nella variabile dipendente che può essere spiegata dalla variabile indipendente. Viene calcolato attraverso il rapporto fra la varianza spiegata della variabile indipendente (somma dei quadrati tra i gruppi) e la varianza totale (somma totale dei quadrati)

$$\eta^2 = \frac{SSB}{SST}.$$

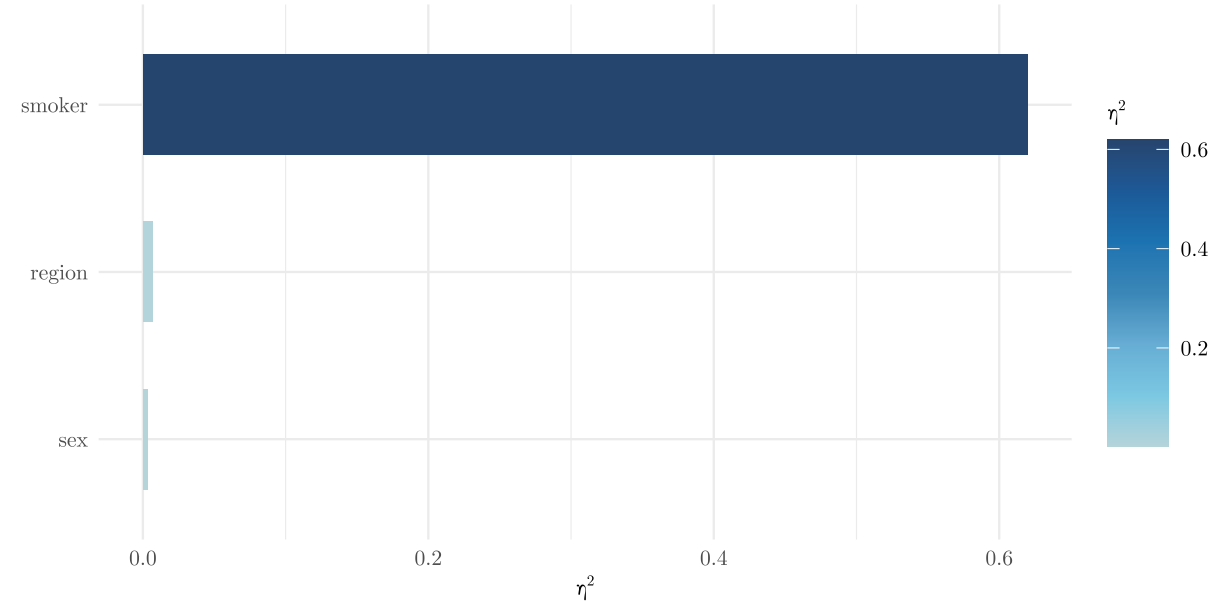
```r
eta2 = function(x,y) {
  m = mean(x,na.rm = TRUE)
  sct = sum((x - m)^2,na.rm = TRUE)
  n = table(y)
  mk = tapply(x,y,mean,na.rm = TRUE)
  sce = sum(n * (mk - m)^2)
  return(ifelse(sct > 0,sce / sct,0))
}

var_qualitative = names(data)[sapply(data,is.factor)]
```

```r
eta2_results = sapply(var_qualitative,function(var) {
  eta2(data$charges,data[[var]])
})
eta2_df = data.frame(variabile = names(eta2_results),eta2 = eta2_results)
eta2_df
```

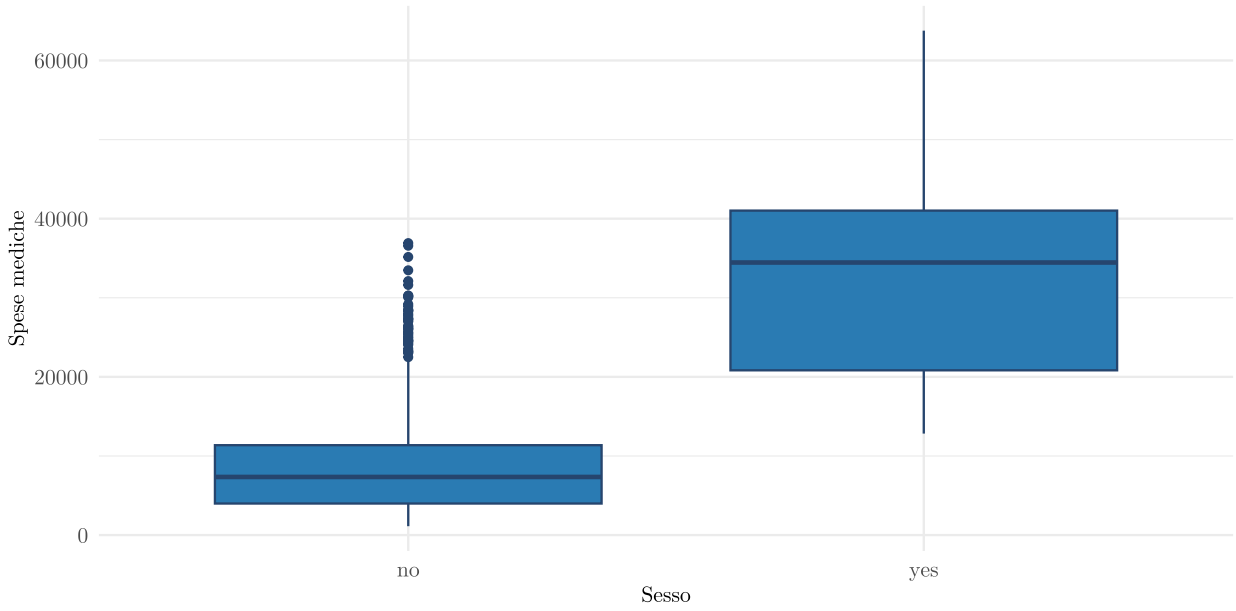
```
##        variabile        eta2
## sex          sex 0.003282380
## smoker    smoker 0.619764815
## region    region 0.006634017
```

```r
library(latex2exp)
eta2_df %>%
  arrange(desc(eta2)) %>%
  ggplot(aes(x = reorder(variabile,eta2),y = eta2,fill = eta2)) +
  geom_bar(stat = "identity",width = 0.6) +
  scale_fill_paletteer_c(`"ggthemes::Classic Blue"`) +
  labs(x = "",y = TeX(sprintf("$\\eta^2$")),fill = TeX(sprintf("$\\eta^2$"))) +
  coord_flip() +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 10),
        legend.title = element_text(size = 10),
        legend.text = element_text(size = 10),
        legend.key.size = unit(1,"cm"))
```



Come ci si può aspettare tra le variabili categoriali, `smoker` è quella maggiormente associata alle spese mediche. Ci si aspetta infatti che se un individuo è fumatore allora l'ammontare di `charges` sarà superiore. Questo aspetto può essere visualizzato confrontando i boxplot della variabile `charges` per le modalità `yes` e `no` di `smoker`.

```r
ggplot(data,aes(y = charges,x = smoker)) +
  geom_boxplot(fill = paletteer_c("ggthemes::Classic Blue",6)[4],
               col = paletteer_c("ggthemes::Classic Blue",6)[6]) +
  labs(x = "Sesso",y = "Spese mediche") +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 10))
```



Le variabili `region` e `sex` non riportano invece un valore di η^2 elevato, dunque non influenzano molto la variabile di interesse.

Alternativamente, considerato che tutte le variabili possono essere pensate come suddivise in classi, possiamo trasfromarle in modo siano categoriali e calcolare per tutte l'indice η^2. Riportiamo di seguito le classi per la trasformazione delle variabili:

- `age`: adottiamo l'usuale suddivisione 18-24, 25-34, 35-49, 50-64, ottenendo all'interno delle classi una numerosità adegata.

```r
data2 = data
data2$age = cut(data2$age,breaks = c(18,25,35,50,65),
                labels = c("18-24","25-34","35-49","50-64"),
                include.lowest = T,right = F)
table(data2$age)
```

```
##
## 18-24 25-34 35-49 50-64
##   278   271   404   385
```

- `bmi`: rispettiamo la definizione dell'indice di massa corporea definendo le classi sottopeso, normopeso, sovrappeso, obeso ed estremamente obeso.

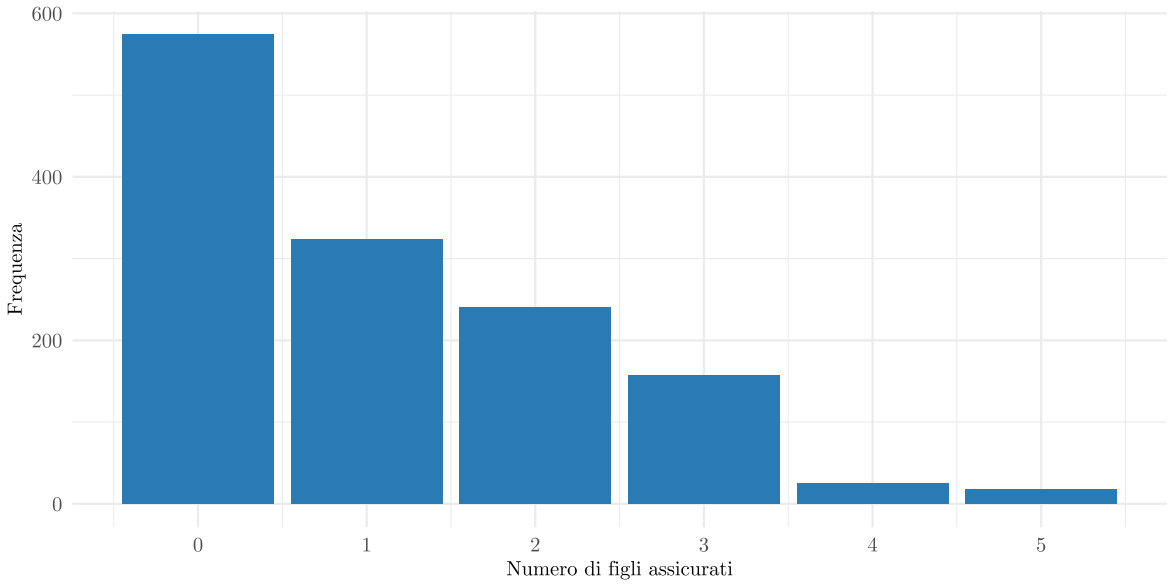
```r
data2$bmi = cut(data2$bmi,breaks = c(0,18.5,24.9,29.9,34.9,100),
                labels = c("Sottopeso","Normopeso","Sovrappeso","Obeso",
                           "Estremamente obeso"),right = F)
table(data2$bmi)
```

```
##
##          Sottopeso          Normopeso         Sovrappeso              Obeso
##                 20                222                377                399
## Estremamente obeso
##                320
```

Osservando la suddivisione in classi di `bmi` si nota che la maggior parte degli individui sono in sovrappeso, obesi o estremamente obesi.

- `children`: per individuare le classi visualizzazione la distribuzione di questa variabile.

```
ggplot(data,aes(x = children)) +
  geom_bar(fill = paletteer_c("ggthemes::Classic Blue",6)[4]) +
  labs(x = "Numero di figli assicurati",y = "Frequenza") +
  scale_x_continuous(breaks = 0:5) +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 10))
```



La numerosità per gli individui che hanno 4 o 5 figli assicurati è ridotta. Possiamo quindi pensare di accorpare queste due modalità con chi ha 3 figli coperti dall'assicurazione. Verranno quindi formate le seguenti classi: 0, 1, 2, 3+.

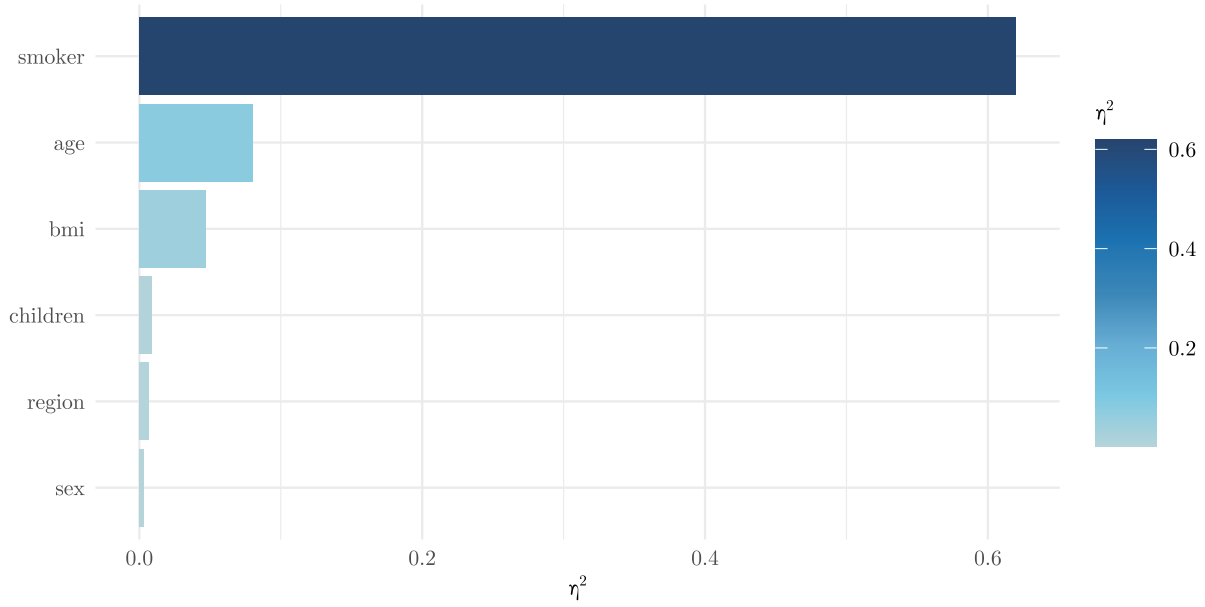
```
data2$children = cut(data2$children,breaks = c(0,1,2,3,6),
                     labels = c("0","1","2","3+"),
                     include.lowest = T,right = F)
table(data2$children)
```

```
##
##   0   1   2  3+
## 574 324 240 200
```

Si nota che ben 898 individui su 1338 (67%) ha al massimo un figlio assicurato.

Calcoliamo ora l'indice di associazione η^2 tra le variabili categoriali e `charges`.

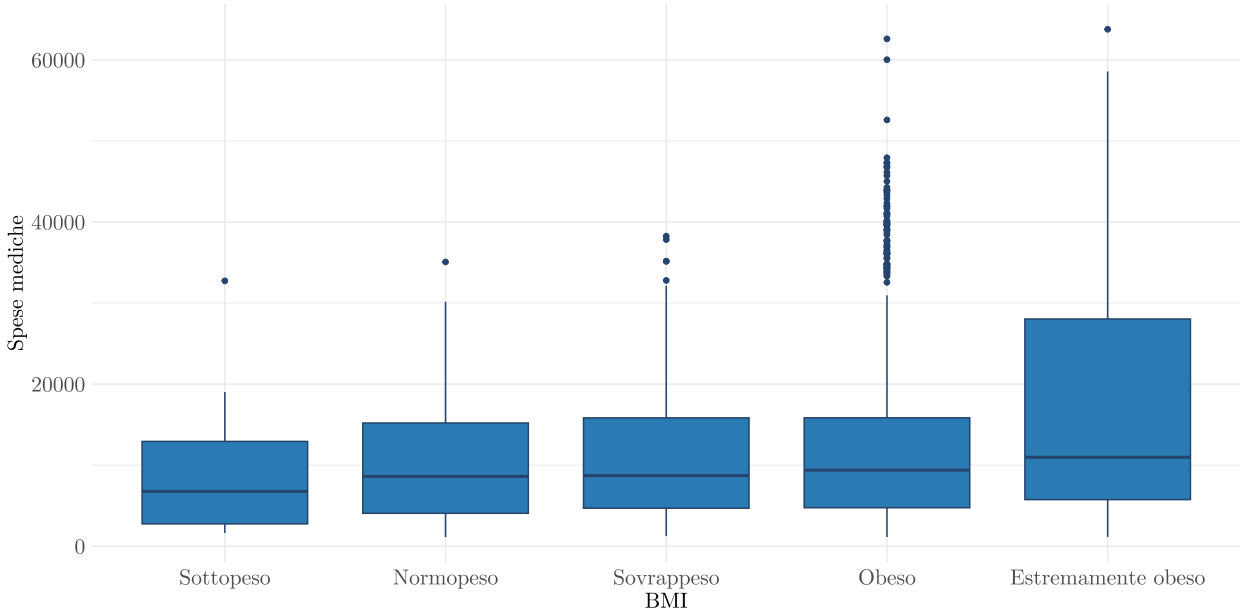
```
eta2_df %>%
  arrange(desc(eta2)) %>%
  ggplot(aes(x = reorder(variabile,eta2),y = eta2,fill = eta2)) +
  geom_bar(stat = "identity") +
  scale_fill_paletteer_c(`"ggthemes::Classic Blue"`) +
  labs(x = "",y = TeX(sprintf("$\\eta^2$")),fill = TeX(sprintf("$\\eta^2$"))) +
  coord_flip() +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 10),
        legend.title = element_text(size = 10),
        legend.text = element_text(size = 10),
        legend.key.size = unit(1,"cm"))
```



La variabile più associata alle spese mediche è `smoker`, seguita da `age` e `bmi`. Il numero di figli coperti da assicurazione, la regione di provenienza e il sesso degli individui sembrano essere poco associate con `charges`.

Visualizziamo anche la distribuzione delle spese mediche per le varie classi di `bmi`.

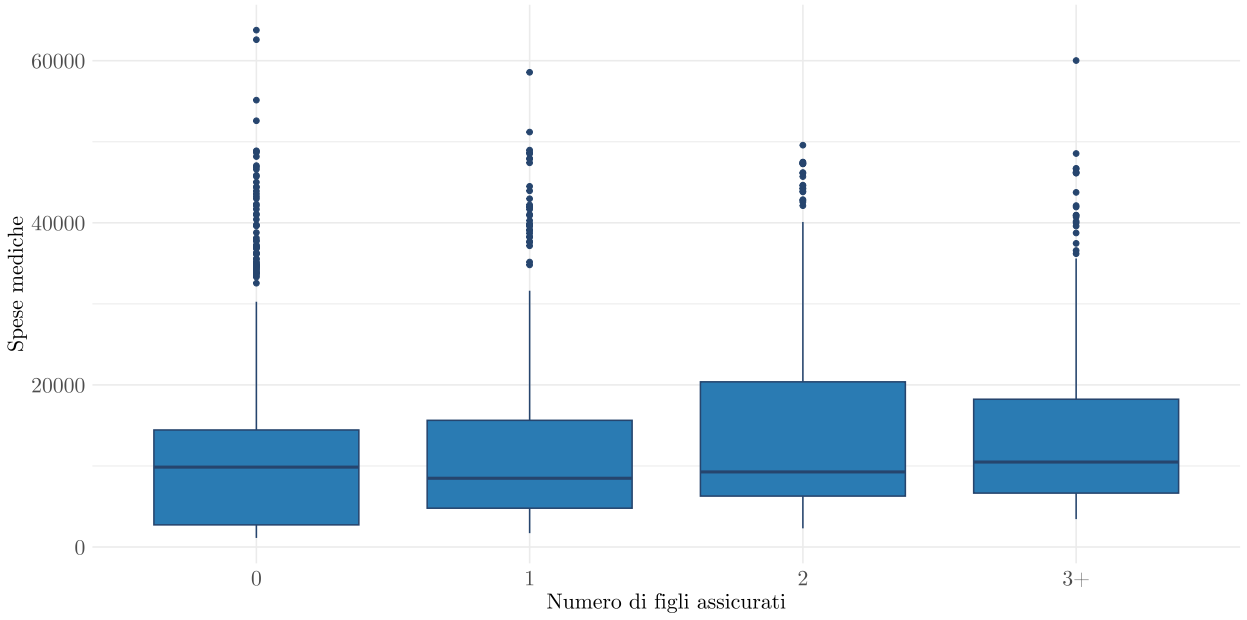
```
ggplot(data2,aes(y = charges,x = bmi)) +
  geom_boxplot(fill = paletteer_c("ggthemes::Classic Blue",6)[4],
               col = paletteer_c("ggthemes::Classic Blue",6)[6]) +
  labs(x = "BMI",y = "Spese mediche") +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 15),
        axis.title = element_text(size = 15))
```



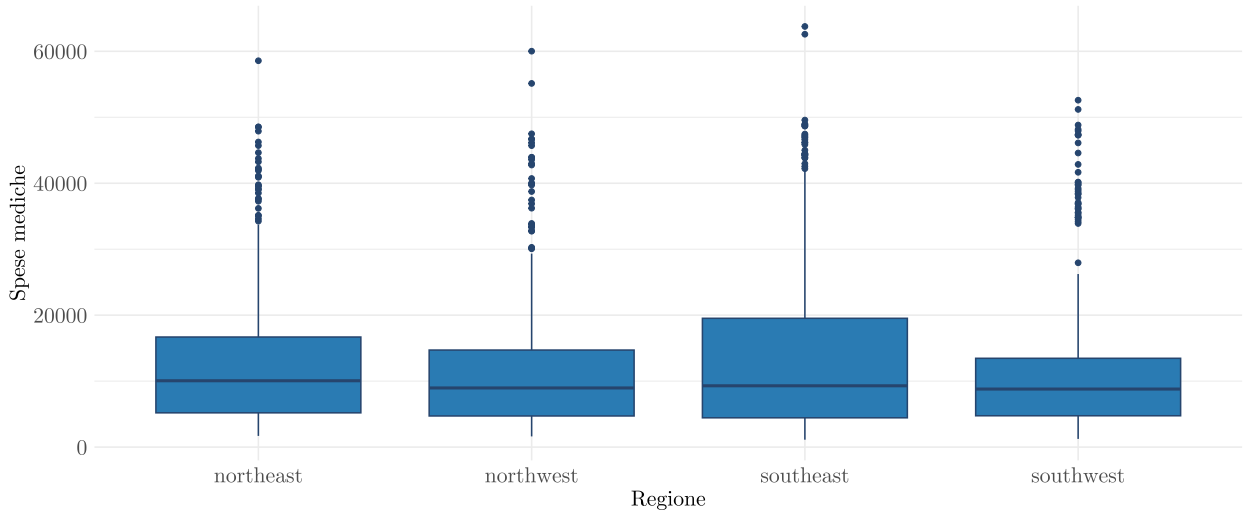
All'aumentare delle classi di `bmi` si osserva un leggero aumento delle spese mediche.

Per completezza riportiamo anche i boxplot considerando le variabili meno associate con `charges`.

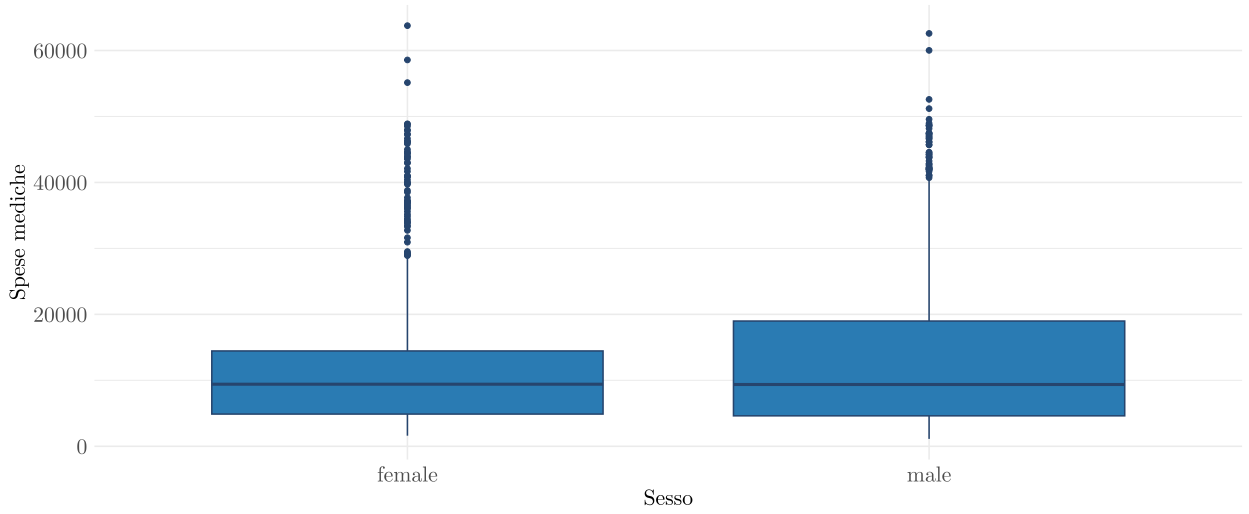
```
ggplot(data2,aes(y = charges,x = children)) +
  geom_boxplot(fill = paletteer_c("ggthemes::Classic Blue",6)[4],
               col = paletteer_c("ggthemes::Classic Blue",6)[6]) +
  labs(x = "Numero di figli assicurati",y = "Spese mediche") +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 15),
        axis.title = element_text(size = 15))
```



```r
ggplot(data2,aes(y = charges,x = region)) +
  geom_boxplot(fill = paletteer_c("ggthemes::Classic Blue",6)[4],
               col = paletteer_c("ggthemes::Classic Blue",6)[6]) +
  labs(x = "Regione",y = "Spese mediche") +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 15),
        axis.title = element_text(size = 15))
```




```r
ggplot(data,aes(y = charges,x = sex)) +
  geom_boxplot(fill = paletteer_c("ggthemes::Classic Blue",6)[4],
               col = paletteer_c("ggthemes::Classic Blue",6)[6]) +
  labs(x = "Sesso",y = "Spese mediche") +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 15),
        axis.title = element_text(size = 15))
```



Regressione lineare

Un'alternativa per vedere come le variabili influenzano `charges` è stimatare un modello di regressione lineare e valutare i coefficienti ottenuti.

```
lm0 = lm(charges ~ .,data = data)
summary(lm0)
```

```
##
## Call:
## lm(formula = charges ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Un semplice modello lineare ci fornisce informazioni leggermente differenti rispetto a quanto ottenuto in precedenza. Oltre ad `age`, `bmi` e `smoker`, secondo il modello anche il numero di figli coperti dall'assicurazione ha un effetto significativo sulle spese mediche e il coefficiente ha segno positivo. Come emerso anche dai boxplot, fermo restando le altre variabili, al crescere del numero di figli assicurati anche le spese mediche aumentano. I parametri relativi alla regione, specificatamente quelli relativi al sud, risultano significativi anche se non fortemente. Possiamo quindi affermare che vi è una leggera differenza tra nord e sud considerando le spese mediche. Inoltre, vale la pena porre l'attenzione sulla grandezza assunta dal coefficiente relativo all'essere o meno un fumatore, confermando nuovamente l'elevata importanza di questa variabile.

Considerata l'assimetria della variabile `charges` possiamo pensare di utilizzare la trasformazione logaritmica.

```
lm0.log = lm(log(charges) ~ .,data = data)
summary(lm0.log)
```

```
##
## Call:
## lm(formula = log(charges) ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07186 -0.19835 -0.04917  0.06598  2.16636
##
## Coefficients:
```

```
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.0305581  0.0723960  97.112  < 2e-16 ***
## age              0.0345816  0.0008721  39.655  < 2e-16 ***
## sexmale         -0.0754164  0.0244012  -3.091 0.002038 **
## bmi              0.0133748  0.0020960   6.381 2.42e-10 ***
## children         0.1018568  0.0100995  10.085  < 2e-16 ***
## smokeryes        1.5543228  0.0302795  51.333  < 2e-16 ***
## regionnorthwest -0.0637876  0.0349057  -1.827 0.067860 .
## regionsoutheast -0.1571967  0.0350828  -4.481 8.08e-06 ***
## regionsouthwest -0.1289522  0.0350271  -3.681 0.000241 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4443 on 1329 degrees of freedom
## Multiple R-squared:  0.7679, Adjusted R-squared:  0.7666
## F-statistic: 549.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Ora tutte le variabili risultano significative. Il valore del coefficiente di determinazione R^2 non ha però subito un grande miglioramento.

Model-based clustering

Potrebbe essere di interesse valutare l'esistenza di gruppi di assicurati con caratteristiche simili in termini dell'ammontare delle spese mediche addebitate alla compagnia assicurativa, valutando poi l'impatto delle altre variabili disponibili su questi gruppi e, di conseguenza, sulla grandezza della variabile `charges`.

Per farlo utilizziamo la metodologia di *clustering* basata su modello (*model-based clustering*). Questa prevede che le osservazioni siano realizzazioni di una combinazione pesata di distribuzioni di probabilità, chiamata **mistura di distribuzioni**. Ogni componente della mistura rappresenterà un singolo gruppo.

Mistura di distribuzioni

Definiamo ora una mistura finita di distribuzioni. Sia $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ un campione di n osservazioni indipendenti ed identicamente distribuite. Supponendo di suddividere la popolazione in G *cluster*, una mistura di distribuzioni è definita come

$$f(\mathbf{x}_i|\boldsymbol{\Psi}) = \sum_{k=1}^{G} \pi_k f_k(x_i; \boldsymbol{\theta}_k), \tag{1}$$

dove $\boldsymbol{\Psi} = \{\pi_1, \ldots, \pi_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G\}$ sono i parametri della mistura, $f_k(x_i|\boldsymbol{\theta}_k)$ è la densità della k-esima componente caratterizzata da un vettore di parametri $\boldsymbol{\theta}_k$, $(\pi_1, \ldots, \pi_G)$ sono chiamate *mixing probabilities* e indicano la probabilità di appartenere ai singoli gruppi e sono tali che

$$\sum_{k=1}^{G} \pi_k = 1, \quad 0 \leq \pi_k \leq 1, \forall k = 1, \ldots, G,$$

e G è il numero di componenti della mistura o, equivalentemente, il numero di *cluster*.

Nella maggior parte delle applicazioni le componenti provengono da una stessa famiglia di distribuzioni, come ad esempio la Gaussiana. Questa scelta porta ai **modelli di mistura Gaussiana** che assume quindi una distribuzione normale per ogni componente della mistura, $f_k(x; \boldsymbol{\theta}_k) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \forall k = 1, \ldots, G$. In questo caso, i gruppi avranno una forma ellissoidale, centrati sulla media $\boldsymbol{\mu}_k$ e con alcune caratteristiche quali forma, dimensione e orientamento determinati dalla forma della matrice di varianza e covarianza $\boldsymbol{\Sigma}_k$.

Un criterio per controllare le caratteristiche dei *cluster* è quello di effettuare una decomposizione a valori singolari della matrice di varianze e covarianze

$$\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^{'}$$

dove $\mathbf{D}_k$ è la matrice ortogonale degli autovettori di $\mathbf{\Sigma}_k$, $\lambda_k = |\mathbf{\Sigma}_k|^{\frac{1}{p}}$ e $\mathbf{A}_k$ consiste in una matrice diagonale tale che $|\mathbf{A}_k| = 1$ i cui elementi sulla diagonale corrispondono agli autovalori normalizzati di $\mathbf{\Sigma}_k$ disposti in ordine decrescente. Quindi, $\mathbf{D}_k$ regola l'orientamento della k-esima componente della mistura, λ_k governa la sua dimensione mentre $\mathbf{A}_k$ controlla la sua forma. Queste caratteristiche geometriche della distribuzione sono stimate dai dati e possono variare tra *cluster* oppure rimanere le stesse per tutti i gruppi. Nella tabella 1 vengono riportati i 14 possibili modelli con diverse caratteristiche geometriche che possono essere specificati.

Modello	$\mathbf{\Sigma}_k$	Orientamento	Volume	Forma
`EII`	$\lambda\mathbf{I}$	-	Uguale	Sferica
`VII`	$\lambda_k\mathbf{I}$	-	Variabile	Sferica
`EEI`	$\lambda\mathbf{A}$	Allineato con gli assi	Uguale	Uguale
`VEI`	$\lambda_k\mathbf{A}$	Allineato con gli assi	Variabile	Uguale
`EVI`	$\lambda\mathbf{A}_k$	Allineato con gli assi	Uguale	Variabile
`VVI`	$\lambda_k\mathbf{A}_k$	Allineato con gli assi	Variabile	Variabile
`EEE`	$\lambda\mathbf{DAD}'$	Uguale	Uguale	Uguale
`VEE`	$\lambda_k\mathbf{DAD}'$	Uguale	Variabile	Uguale
`EVE`	$\lambda\mathbf{DA}_k\mathbf{D}'$	Uguale	Uguale	Variabile
`EEV`	$\lambda\mathbf{D}_k\mathbf{AD}_k^{'}$	Variabile	Uguale	Uguale
`VVE`	$\lambda_k\mathbf{DA}_k\mathbf{D}'$	Uguale	Variabile	Variabile
`VEV`	$\lambda_k\mathbf{D}_k\mathbf{AD}_k^{'}$	Variabile	Variabile	Uguale
`EVV`	$\lambda\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^{'}$	Variabile	Uguale	Variabile
`VVV`	$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^{'}$	Variabile	Variabile	Variabile

Table 1: Modelli di mistura Gaussiani

Si noti che nel caso unidimensionale si possono avere al massimo due modelli: `E` che prevede un uguale varianza tra le componenti e `V` che invece assume che i gruppi presentino varianze differenti.

I parametri del modello, $\mathbf{\Psi}$, sono ignoti e dovranno quindi essere stimati. Stimare direttamente la funzione di log-verosimiglianza di 1, risulta complesso e per questo la stima di massima verosimiglianza viene ottenuta tramite l'**algoritmo EM**.

Infine, come criterio per la selezione del miglior modello viene solitamente utilizzato il BIC.

Applicazione al dataset `insurance`

Per effettuare *model-based clustering* sull'insieme di dati relativo alle spese mediche utilizziamo il pacchetto `mclust` e consideriamo solamente la variabile `charges`.

```
set.seed(123)
library(mclust)

mbc = Mclust(data$charges,verbose = F)
summary(mbc)
```

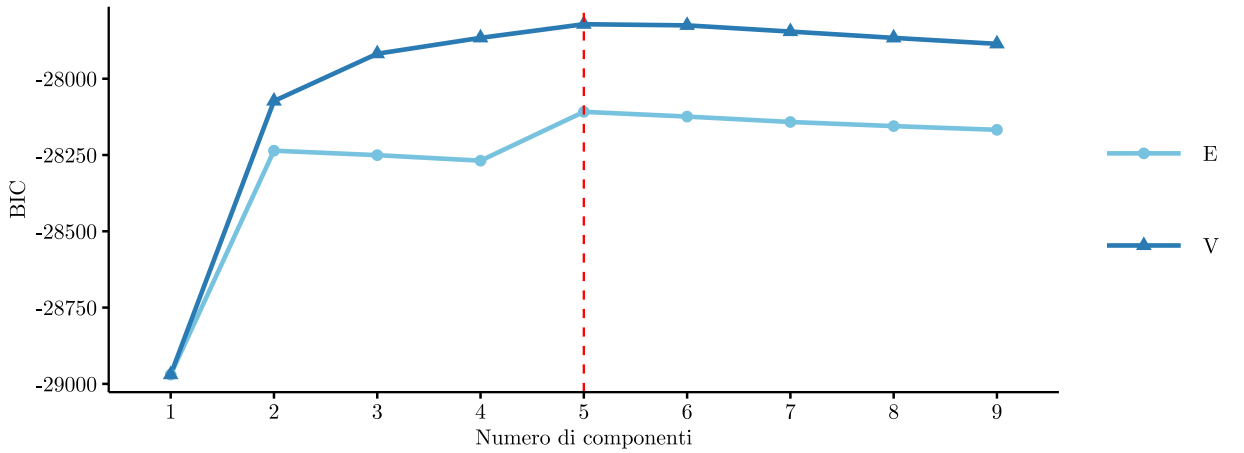
```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust V (univariate, unequal variance) model with 5 components:
##
```

```
##  log-likelihood    n df      BIC       ICL
##       -13860.36 1338 14 -27821.5 -28305.71
##
## Clustering table:
##   1   2   3   4   5
## 173 327 478 191 169
```

Il comando `Mclust` ha stimato solamente due modelli (`E` e `V`) avendo considerato solamente le spese mediche e ha considerato un numero di componenti che vanno da 1 a 9 (di *default*). Dal *summary* del modello si nota che è selezionato un modello con $G = 5$ componenti che prevede diverse varianze tra i *cluster*. Si può inoltre vedere che il *cluster* 3 risulta quello più numeroso, seguito dal secondo, mentre i restanti tre presentano una numerosità più ridotta.

Possiamo rappresentare l'andamento del BIC per i modelli stimati.

```
library(factoextra)
fviz_mclust_bic(mbc,legend = "right",shape = "model",size = 1,
                palette = paletteer_c("ggthemes::Classic Blue",6)[c(2,4)]) +
  labs(x = "Numero di componenti") +
  theme(legend.title = element_blank(),
        legend.text = element_text(size = 10),
        legend.key.size = unit(1.5,'cm'),
        axis.title = element_text(size = 10),
        axis.text = element_text(size = 10),
        title = element_blank(),
        text = element_text(family = "CMUSerif"))
```



Possiamo anche considerare altri criteri per la selezione del modello, come l'AIC o l'ICL (*Integrated Complete Likelihood*). Considerato il grafico precedente, in cui si nota che il modello con uguali varianze per i gruppi non risulta preferibile per nessun numero di componenti, stimiamo solamente il modello `V`.

```
AIC = c()
for(i in 1:5){
  AIC[i] = AIC(Mclust(data$charges,verbose = F,modelNames = "V",G = i))
}
AIC
```

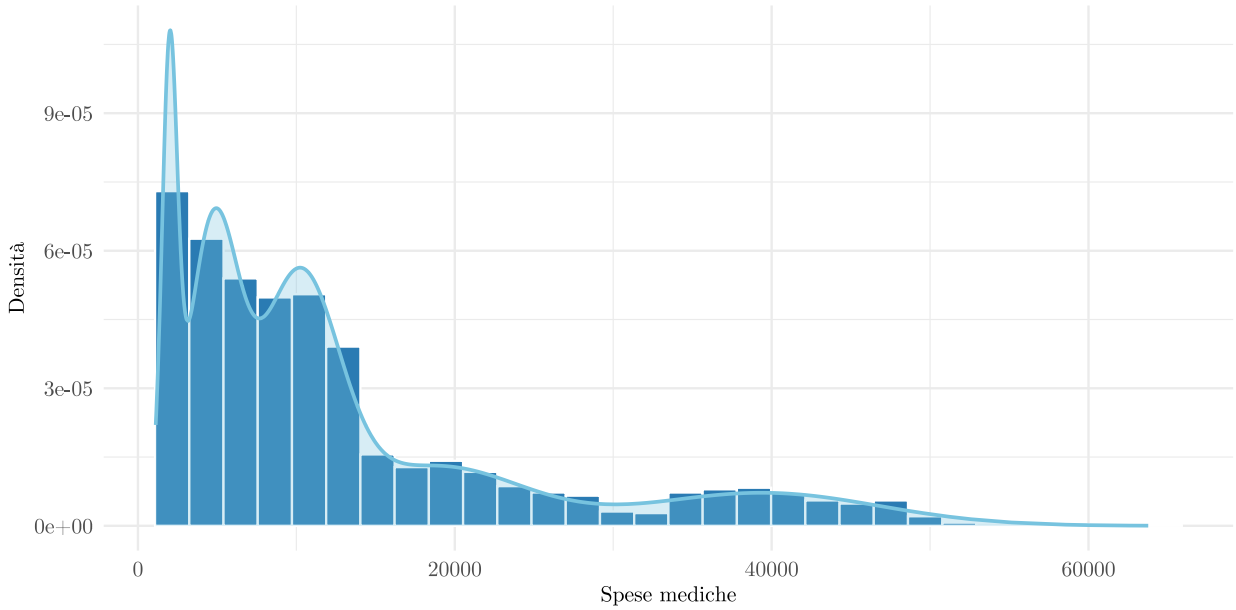
```
## [1] 28959.26 28047.20 27876.51 27808.96 27748.72
```

Anche l'AIC porta a selezionare un modello `V` con 5 componenti.

Rappresentiamo la mistura stimata per la variabile `charges`.

```
df = data.frame(charges = data$charges,
                density = mbc[["parameters"]][["pro"]][1] *
                  dnorm(data$charges,mean = mbc[["parameters"]][["mean"]][["1"]],
                        sd = sqrt(mbc[["parameters"]][["variance"]]$sigmasq[1])) +
                  mbc[["parameters"]][["pro"]][2] *
                  dnorm(data$charges,mean = mbc[["parameters"]][["mean"]][["2"]],
                       sd = sqrt(mbc[["parameters"]][["variance"]]$sigmasq[2])) +
                  mbc[["parameters"]][["pro"]][3] *
                  dnorm(data$charges,mean = mbc[["parameters"]][["mean"]][["3"]],
                       sd = sqrt(mbc[["parameters"]][["variance"]]$sigmasq[3])) +
                  mbc[["parameters"]][["pro"]][4] *
                  dnorm(data$charges,mean = mbc[["parameters"]][["mean"]][["4"]],
                       sd = sqrt(mbc[["parameters"]][["variance"]]$sigmasq[4])) +
                  mbc[["parameters"]][["pro"]][5]*
                  dnorm(data$charges,mean = mbc[["parameters"]][["mean"]][["5"]],
                       sd = sqrt(mbc[["parameters"]][["variance"]]$sigmasq[5])))

ggplot(df,aes(x = charges,y = density)) +
  geom_histogram(aes(y = ..density..),col = "white",bins = 30,
                 fill = paletteer_c("ggthemes::Classic Blue",6)[4]) +
  geom_line(col = paletteer_c("ggthemes::Classic Blue",6)[2],size = 0.8) +
  geom_area(fill = paletteer_c("ggthemes::Classic Blue",6)[2],alpha = 0.3) +
  labs(x = "Spese mediche",y = "Densità") +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 10))
```



Possiamo notare che la mistura a 5 componenti stimata riesce a cogliere bene la coda della distribuzione. L'adattamento ai valori osservati è peggiore invece per i valori di `charges` inferiori a 20.000.

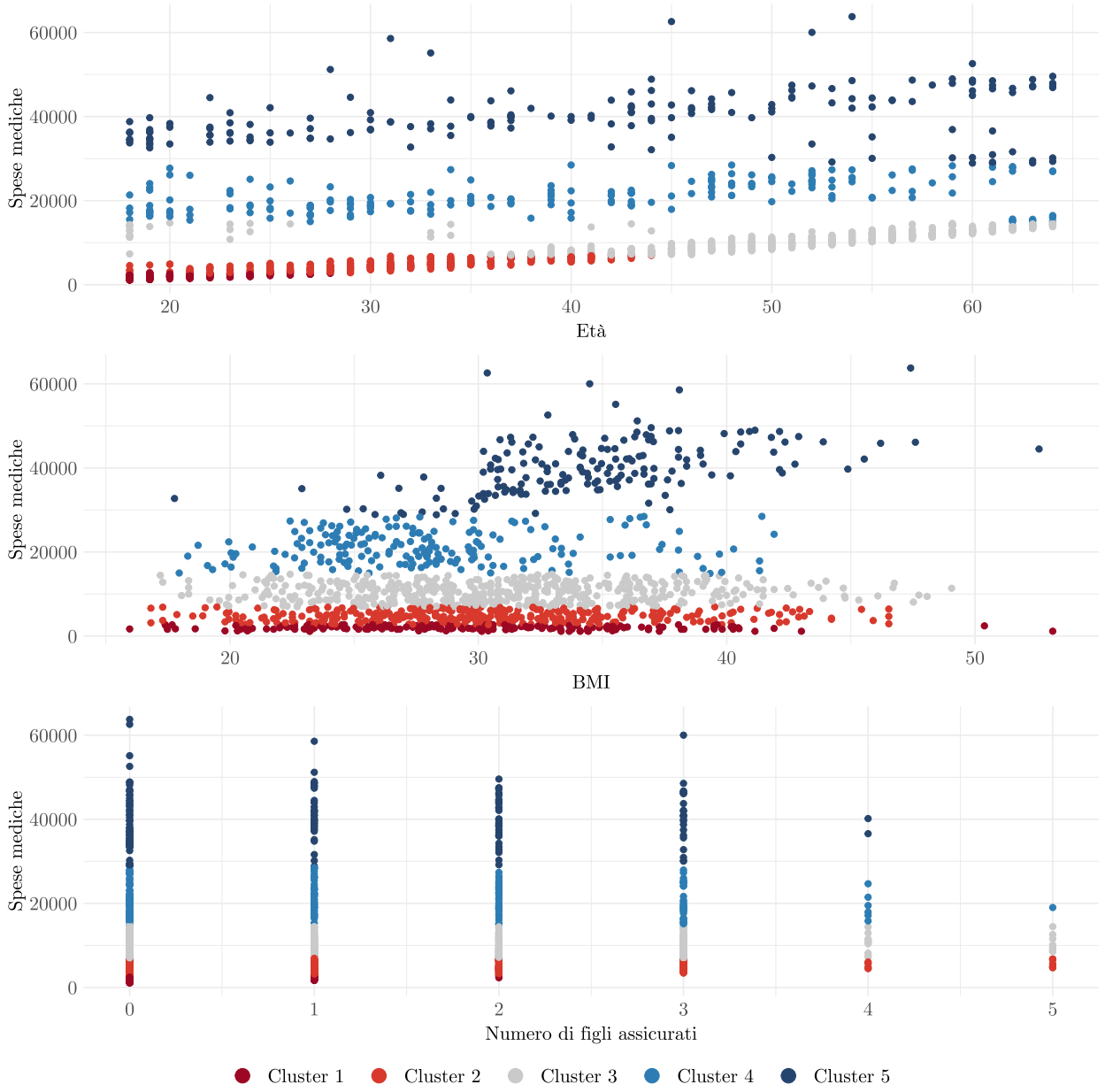
Vediamo anche come sono state suddivise le osservazioni nei 5 *cluster* considerando le variabili `age`, `bmi` e `children`.

```
g1 = ggplot(data,aes(x = age,y = charges,col = factor(mbc[["classification"]]))) +
  geom_point(size = 2) +
  labs(x = "Età",y = "Spese mediche") +
  scale_color_manual(values = paletteer_c("ggthemes::Classic Red-Blue",5),
                     label = paste("Cluster ",seq(1,5,1),sep = "")) +
  labs(col = " ") +
  guides(colour = guide_legend(override.aes = list(size = 5))) +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 15),
        axis.title = element_text(size = 15),
        legend.text = element_text(size = 15),
        legend.title = element_text(size = 15),
        legend.key.size = unit(1,"cm"))

g2 = ggplot(data,aes(x = bmi,y = charges,col = factor(mbc[["classification"]]))) +
  geom_point(size = 2) +
  labs(x = "BMI",y = "Spese mediche") +
  scale_color_manual(values = paletteer_c("ggthemes::Classic Red-Blue",5),
                     label = paste("Cluster ",seq(1,5,1),sep = "")) +
  labs(col = " ") +
  guides(colour = guide_legend(override.aes = list(size = 5))) +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 15),
        axis.title = element_text(size = 15),
        legend.text = element_text(size = 15),
        legend.title = element_text(size = 15),
        legend.key.size = unit(1,"cm"))

g3 = ggplot(data,aes(x = children,y = charges,col = factor(mbc[["classification"]]))) +
  geom_point(size = 2) +
  labs(x = "Numero di figli assicurati",y = "Spese mediche") +
  scale_color_manual(values = paletteer_c("ggthemes::Classic Red-Blue",5),
                     label = paste("Cluster ",seq(1,5,1),sep = "")) +
  labs(col = " ") +
  guides(colour = guide_legend(override.aes = list(size = 5))) +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 15),
        axis.title = element_text(size = 15),
        legend.text = element_text(size = 15),
        legend.title = element_text(size = 15),
        legend.key.size = unit(1,"cm"))

library(ggpubr)
ggarrange(g1,g2,g3,ncol = 1,common.legend = T,legend = "bottom")
```



Dai grafici emerge una chiara suddivisione delle unità per fasce di costo. In particolare, per quanto riguarda l'età, si osserva che i primi due gruppi comprendono individui più giovani, mentre il terzo gruppo raccoglie persone di età maggiore. Gli ultimi due gruppi, invece, includono individui di età variegata in modo uniforme.

Non si nota invece una chiara suddivisione degli individui per valore di BMI o per numero di figli assicurati.

Significatività delle variabili nei *cluster*

Così come abbiamo verificato quali variabili presentassero un impatto su `charges` considerando tutte le unità, possiamo valutare se le stesse variabili risultano significative nei *cluster* individuati. Ricordiamo che nel primo caso solamente le variabili `age`, `bmi` e `children` risultavano significative.

Stimiamo allora 5 modelli lineari uno per ogni gruppo e per farlo sarà innanzitutto necessario salvare l'appartenenza degli individui nei vari gruppi.

```
data$latent_class = mbc[["classification"]]
```

```
lm1 = lm(charges ~ age + sex + bmi + children + region,
         data = data %>%
           filter(latent_class == 1))
summary(lm1)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + region, data = data %>%
##     filter(latent_class == 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.522 -17.510   3.874  11.542  97.736
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     -478.0363    16.1516  -29.597  < 2e-16 ***
## age              145.8681     0.6924  210.674  < 2e-16 ***
## sexmale         -485.4506     3.4290 -141.573  < 2e-16 ***
## bmi                1.2503     0.2773    4.509 1.23e-05 ***
## children         586.1106     4.1420  141.503  < 2e-16 ***
## regionnorthwest -209.6163     5.1696  -40.548  < 2e-16 ***
## regionsoutheast -576.6479     5.1088 -112.873  < 2e-16 ***
## regionsouthwest -593.0749     5.1318 -115.568  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.32 on 165 degrees of freedom
## Multiple R-squared:  0.9977, Adjusted R-squared:  0.9976
## F-statistic: 1.032e+04 on 7 and 165 DF,  p-value: < 2.2e-16
```

Per il primo gruppo tutte le variabili sono significative. Si specifica inoltre che per questo gruppo, così come nel secondo non è stata inclusa la variabile `smoker` in quanto all'interno dei primi due *cluster* sono presenti solamente individui non fumatori e dunque la variabile presenterebbe una sola modalità.

```
table(data %>%
        filter(latent_class == 1) %>%
        select(smoker))
```

```
## smoker
##  no yes
## 173   0
```

```
table(data %>%
        filter(latent_class == 2) %>%
        select(smoker))
```

```
## smoker
##  no yes
## 327   0
```

Procediamo con il secondo gruppo.

```
lm2 = lm(charges ~ age + sex + bmi + children + region,
         data = data %>%
           filter(latent_class == 2))
summary(lm2)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + region, data = data %>%
##     filter(latent_class == 2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -141.22  -98.74  -49.03   55.60  490.71
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2096.9971    56.9153 -36.844   <2e-16 ***
## age               210.4775     1.3773 152.820   <2e-16 ***
## sexmale          -470.1962    15.6694 -30.007   <2e-16 ***
## bmi                 0.9286     1.2900   0.720    0.472
## children          591.3209     6.6552  88.850   <2e-16 ***
## regionnorthwest  -210.8132    21.2181  -9.936   <2e-16 ***
## regionsoutheast  -593.5183    22.7500 -26.089   <2e-16 ***
## regionsouthwest  -582.5045    22.1182 -26.336   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 133.8 on 319 degrees of freedom
## Multiple R-squared:  0.9876, Adjusted R-squared:  0.9874
## F-statistic:  3642 on 7 and 319 DF,  p-value: < 2.2e-16
```

Per il secondo *cluster* solamente la variabile `bmi` risulta essere non significativa.

```
lm3 = lm(charges ~ age + sex + bmi + children + smoker + region,
         data = data %>%
           filter(latent_class == 3))
summary(lm3)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker +
##     region, data = data %>% filter(latent_class == 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2240.9 -1062.1  -219.7   691.4  9307.6
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1687.476    634.830   2.658  0.00813 **
## age              172.719      9.519  18.144  < 2e-16 ***
## sexmale         -337.880    147.484  -2.291  0.02241 *
## bmi               10.626     12.985   0.818  0.41361
## children          96.092     62.223   1.544  0.12318
```

```
## smokeryes       8784.170    690.688  12.718  < 2e-16 ***
## regionnorthwest -246.934    206.605  -1.195  0.23261
## regionsoutheast -377.691    217.917  -1.733  0.08372 .
## regionsouthwest -529.945    203.038  -2.610  0.00934 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1598 on 469 degrees of freedom
## Multiple R-squared:  0.4609, Adjusted R-squared:  0.4517
## F-statistic: 50.12 on 8 and 469 DF,  p-value: < 2.2e-16
```

Nel terzo gruppo, oltre a `bmi`, perdono la significatività anche il numero di figli coperti da assicurazione e la regione. Anche il coefficiente relativo al sesso degli individui risulta poco significativo.

```
lm4 = lm(charges ~ age + sex + bmi + children + smoker + region,
         data = data %>%
           filter(latent_class == 4))
summary(lm4)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker +
##     region, data = data %>% filter(latent_class == 4))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8808.1 -1735.2  -392.2  2339.9  8442.4
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     19905.75    2066.07   9.635  < 2e-16 ***
## age               102.97      18.52   5.560 9.48e-08 ***
## sexmale           139.48     504.58   0.276   0.7825
## bmi               -80.49      61.20  -1.315   0.1901
## children         -276.17     204.92  -1.348   0.1794
## smokeryes       -1075.03     632.14  -1.701   0.0907 .
## regionnorthwest   509.09     688.47   0.739   0.4606
## regionsoutheast   295.40     662.06   0.446   0.6560
## regionsouthwest  -226.05     746.34  -0.303   0.7623
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3411 on 182 degrees of freedom
## Multiple R-squared:  0.175,  Adjusted R-squared:  0.1387
## F-statistic: 4.826 on 8 and 182 DF,  p-value: 2.097e-05
```

Per il quarto gruppo, rimane fortemente significativa solamente la variabile `age`.

```
lm5 = lm(charges ~ age + sex + bmi + children + smoker + region,
         data = data %>%
           filter(latent_class == 5))
summary(lm5)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker +
```

```
##     region, data = data %>% filter(latent_class == 5))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12264.1  -1403.1   -441.8    884.3  24006.1
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -388.87    3298.52  -0.118    0.906
## age               212.53      24.96   8.515 1.15e-14 ***
## sexmale           -86.00     729.78  -0.118    0.906
## bmi               645.06      80.76   7.987 2.55e-13 ***
## children          196.73     318.59   0.618    0.538
## smokeryes       10035.99    1458.35   6.882 1.27e-10 ***
## regionnorthwest    42.20    1094.58   0.039    0.969
## regionsoutheast  -122.19     989.44  -0.123    0.902
## regionsouthwest   690.19    1057.90   0.652    0.515
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4522 on 160 degrees of freedom
## Multiple R-squared:  0.5345, Adjusted R-squared:  0.5112
## F-statistic: 22.96 on 8 and 160 DF,  p-value: < 2.2e-16
```

Infine, nell'ultimo *cluster* oltre alla variabile `age`, ritornano significative anche `bmi` e `smoker`.

Possiamo visualizzare in un unico grafico queste informazioni in modo da comprendere più rapidamente la significatività dei parametri e l'effetto delle variabili su `charges`. Rappresentiamo quindi per ogni modello il coefficiente con relativo intervallo di confidenza al 95%. Per farlo, innanzitutto, creiamo una lista contenente i coefficienti per i modelli stimati per ogni *cluster* (escludendo l'intercetta), il loro intervallo di confidenza e il *p-value*.

```
coef.data_list = lapply(1:5,function(i){
  coef.data = data.frame(var = names(coef(get(paste0("lm",i)))[-1]),
                         coef = as.vector(coef(get(paste0("lm",i)))[-1]),
                         min_ci = as.vector(confint(get(paste0("lm",i)),
                                                    level = 0.95)[-1,1]),
                         max_ci = as.vector(confint(get(paste0("lm",i)),
                                                    level = 0.95)[-1,2]),
                         sign = as.vector(summary(get(paste0("lm",i)))$coef[-1,4]))
  coef.data$sign = round(coef.data$sign,4)
  return(coef.data)
})
```

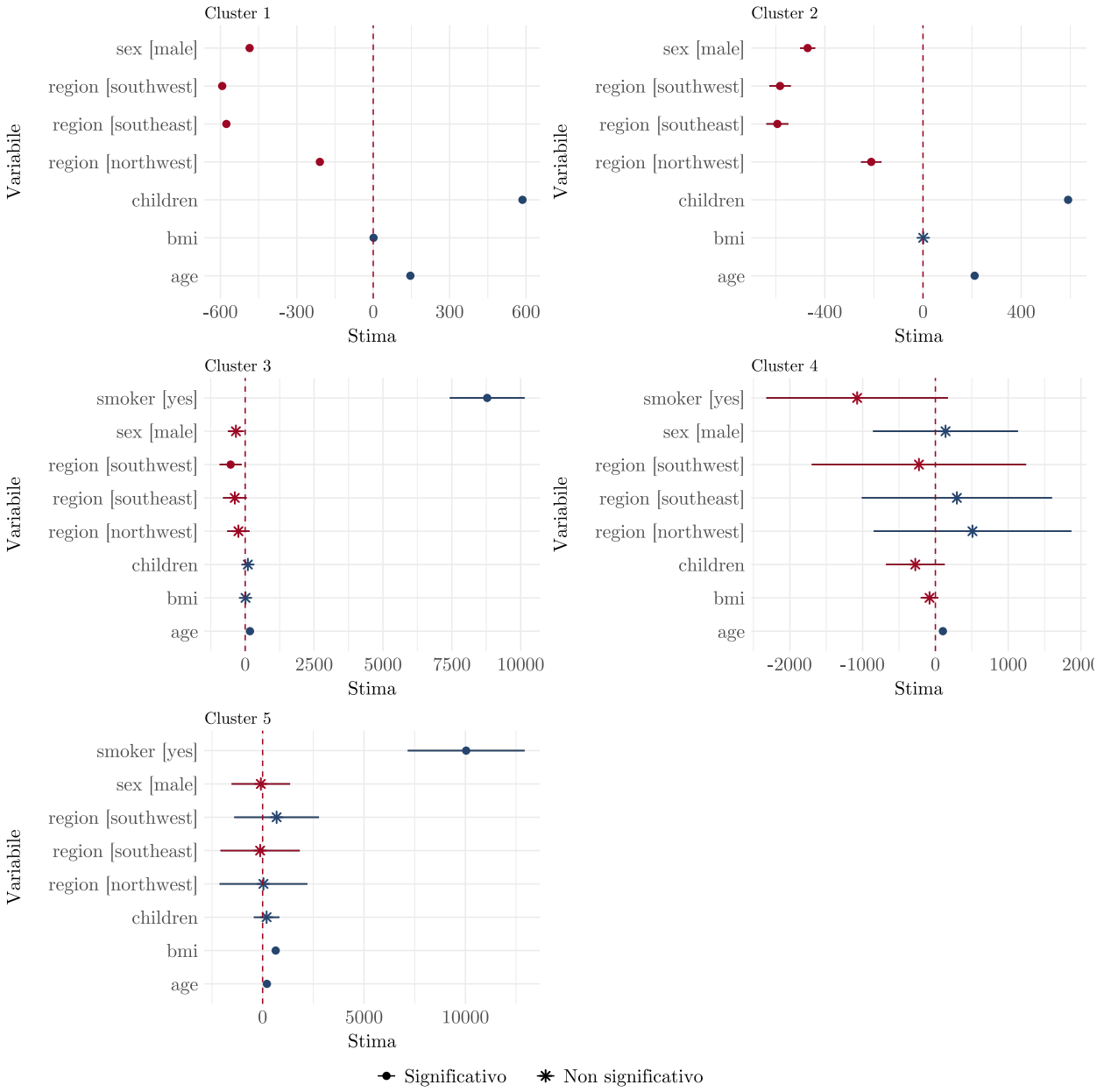
Modifichiamo il nome delle variabili categoriali per rendere più leggibile il grafico.

```
coef.data_list = lapply(coef.data_list,function(coef.data){
  coef.data$var = gsub("regionnorthwest","region [northwest]",coef.data$var)
  coef.data$var = gsub("regionsouthwest","region [southwest]",coef.data$var)
  coef.data$var = gsub("regionnortheast","region [northeast]",coef.data$var)
  coef.data$var = gsub("regionsoutheast","region [southeast]",coef.data$var)
  coef.data$var = gsub("sexmale","sex [male]",coef.data$var)
  coef.data$var = gsub("smokeryes","smoker [yes]",coef.data$var)
  return(coef.data)
})
```

Realizziamo ora il grafico.

```
library(forcats)
plot_list = lapply(1:5,function(i) {
  coef.data = coef.data_list[[i]]
  if(i %in% 1:4){
    coef.data %>%
      ggplot(aes(y = var,x = coef,xmin = min_ci,xmax = max_ci)) +
      geom_linerange() +
      geom_pointrange(size = 0.6,
                      col = ifelse(coef.data$coef < 0,"#9C0824FF","#26456EFF"),
                      pch = ifelse(coef.data$sign < 0.01,16,8)) +
      geom_vline(xintercept = 0,col = "#9C0824FF",lty = "dashed") +
      labs(x = "Stima",y = "Variabile",title = paste("Cluster",i,sep = " ")) +
      theme_minimal() +
      theme(text = element_text(family = "CMUSerif"),
            axis.text = element_text(size = 15),
            axis.title = element_text(size = 15),
            legend.text = element_text(size = 15))
  }
  else{
    coef.data %>%
      ggplot(aes(y = var,x = coef,xmin = min_ci,xmax = max_ci,
                 shape = factor(sign > 0.01))) +
      geom_linerange() +
      geom_pointrange(size = 0.6,
                      col = ifelse(coef.data$coef < 0,"#9C0824FF","#26456EFF")) +
      scale_shape_manual(values = ifelse(coef.data$sign < 0.01,16,8),
                         label = ifelse(coef.data$sign < 0.01,
                                        "Significativo","Non significativo")) +
      geom_vline(xintercept = 0,col = "#9C0824FF",lty = "dashed") +
      labs(x = "Stima",y = "Variabile",
           title = paste("Cluster",i,sep = " "),shape = "") +
      theme_minimal() +
      theme(text = element_text(family = "CMUSerif"),
            axis.text = element_text(size = 15),
            axis.title = element_text(size = 15),
            legend.text = element_text(size = 15))
  }
})

ggarrange(plotlist = plot_list,ncol = 2,nrow = 3,
          common.legend = T,legend = "bottom")
```



Si specifica che nel grafico i punti rossi rappresentano i coefficienti con segno negativo mentre quelli blu quelli con segno positivo. È stata poi riportata una linea verticale tratteggiata in corrispondenza dello zero per facilitare lettura del grafico ed evidenziare la significatività dei parametri. Tuttavia, considerata la grandezza (in valore assoluto) di molti coefficienti potrebbe succedere che un parametro sembri in corrispondenza dello zero ma in realtà non lo sia e rimanga quindi significativo (come per `bmi` in `lm0`). Per ovviare a questo problema di lettura del grafico sono stati rappresentati due tipologie di punti differenti, come riportato in legenda: il punto pieno rappresenta i coefficienti significativi, mentre l'asterisco quelli non significativi.

Nota: creando un unico codice per tutti i modelli, la legenda comune si riferirà al modello `lm0`. Quest'ultimo presenta però tutti i coefficienti significativi e ne consegue che la legenda riportata avrà solamente l'etichetta "Significativo". Per questo motivo, è stata inserita la legenda solamente per l'ultimo modello, che avendo sia coefficienti significativi che non significativi, permette di visualizzare la legenda correttamente.

Caratteristiche dei *cluster*

Valutiamo ora le caratteristiche di ogni gruppo considerando le altre variabili presenti nel dataset. Per fare ciò, consideriamo le variabili categoriali contenute in `data2`.

```
data2$latent_class = mbc[["classification"]]
```

Iniziamo considerando la variabile `age`:

- *Cluster* 1: possiamo notare che il primo gruppo è caratterizzato principalmente da individui giovani di età compresa tra 18 e 24 anni, vi è però una piccola percentuale di persone nella fascia di età 25-34. Possiamo comunque affermare che il primo *cluster* è composto principalmente da individui giovani.
- *Cluster* 2: il secondo gruppo è composto principalmente da individui di età compresa tra i 25 e i 49 anni e in misura decisamente inferiore da giovani di 18-24 anni. Questo gruppo include quindi individui di età più avanzata rispetto al primo *cluster*.
- *Cluster* 3: il terzo *cluster* raggruppa individui con una fascia di età 35-64. Ne consegue che questo gruppo è composto da individui di età medio/alta, anche se include alcuni giovani.
- *Cluster* 4 e 5: gli ultimi due gruppi includono invece individui di età variegata, con una distribuzione uniforme tra le fasce di età.

```
table(data2$age,data2$latent_class)
```

```
##
##           1   2   3   4   5
##   18-24 158  37  16  33  34
##   25-34  15 179   5  45  27
##   35-49   0 111 184  60  49
##   50-64   0   0 273  53  59
```

Passiamo alla variabile `sex`: questa variabile risulta significativa solamente per i primi due *cluster*. Tuttavia, in entrambi i casi vi è una buona numerosità di individui di entrambi i sessi. Questo si riscontra anche per i restanti gruppi.

```
table(data2$sex,data2$latent_class)
```

```
##
##            1   2   3   4   5
##   female  71 171 260  95  65
##   male   102 156 218  96 104
```

Consideriamo ora la variabile `bmi`: l'indice di massa corporea è significativo solamente per il primo *cluster* che include maggiormente persone in sovrappeso e obese. Sono però presenti anche 36 individui normopeso e una piccolissima percentuale di persone sottopeso. In generale, si può vedere che la numerosità più elevata si trova in corrispondenza di un BMI più elevato. Questo è probabilmente dovuto alla grande numerosità di individui in quelle fasce di BMI.

```
table(data2$bmi,data2$latent_class)
```

```
##
##                       1   2   3   4   5
##   Sottopeso           5   6   6   2   1
##   Normopeso          36  57  73  54   2
##   Sovrappeso         45 103 132  82  15
##   Obeso              49 100 147  29  74
##   Estremamente obeso 38  61 120  24  77
```

Per quanto riguarda la variabile `children`, questa risulta significativa solamente i primi due *cluster*. Possiamo affermare che il primo gruppo è composto principalmente da individui con nessun figlio assicurato, mentre il secondo gruppo include principalmente individui con al massimo 2 figli coperti da assicurazione. Nei restanti gruppi non si distingono particolari caratteristiche e includono una buona percentuale di individui di tutte le modalità di `children`.

```
table(data2$children,data2$latent_class)
```

```
##
##        1   2   3   4   5
##   0  144  93 198  70  69
##   1   27 102 113  47  35
##   2    2  75  86  37  40
##   3+   0  57  81  37  25
```

Passiamo alla variabile `smoker`. Abbiamo già fatto notare in precedenza che i primi due gruppi includono solamente persone che non fumano. Per quanto riguarda gli altri gruppi, questa variabile non risulta significativa solamente per il terzo *cluster* per il quale si può osservare una maggioranza di individui non fumatori. Gli ultimi due gruppi includono maggiormente persone fumatrici.

```
table(data2$smoker,data2$latent_class)
```

```
##
##         1   2   3   4   5
##   no  173 327 471  81  12
##   yes   0   0   7 110 157
```

Infine, consideriamo la variabile `region`. Questa variabile risulta significativa solamente per i primi due gruppi. Ad ogni modo, in tutti i casi non si riesce a definire una regione più rappresentata rispetto alle altre.

```
table(data2$region,data2$latent_class)
```

```
##
##               1   2   3   4   5
##   northeast  32  75 127  53  37
##   northwest  41  87 116  47  34
##   southeast  56  84 108  55  61
##   southwest  44  81 127  36  37
```

Approfondimento

Invece di utilizzare il BIC per la selezione del miglior modello di mistura utilizziamo l'ICL (*Integrated Complete Likelihood*).

```
set.seed(123)
icl = mclustICL(data$charges,verbose = F)
icl
```

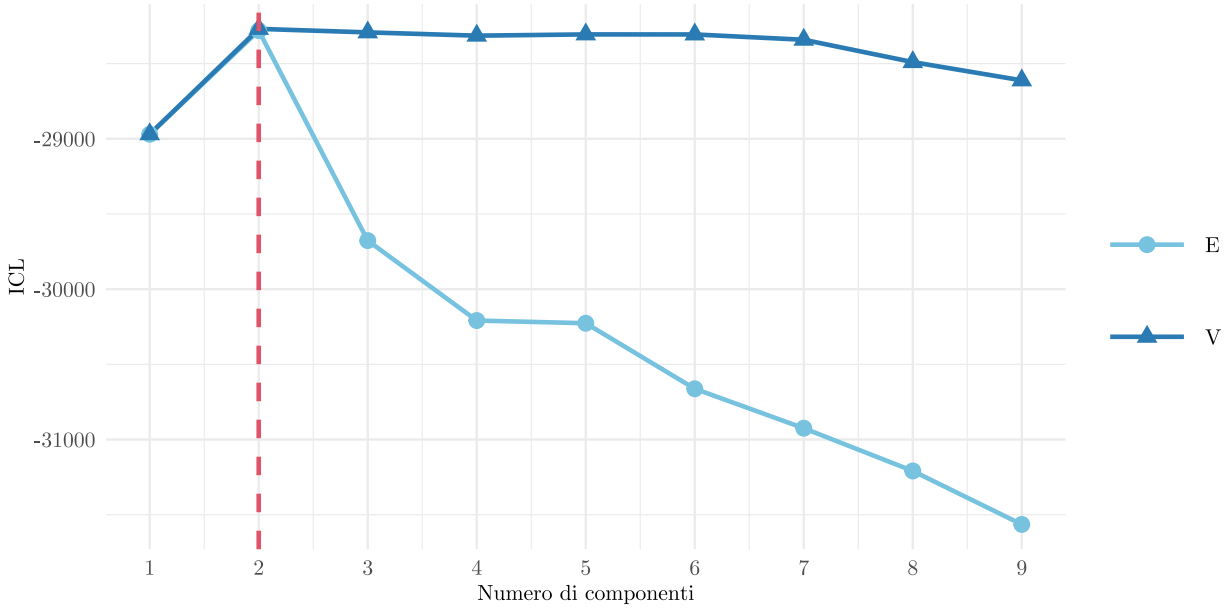
```
## Integrated Complete-data Likelihood (ICL) criterion:
##           E         V
## 1 -28969.66 -28969.66
## 2 -28280.50 -28268.94
## 3 -29677.02 -28292.16
## 4 -30208.74 -28314.19
## 5 -30226.72 -28305.71
## 6 -30662.60 -28306.10
## 7 -30925.03 -28341.30
## 8 -31209.33 -28489.96
```

```
## 9 -31564.85 -28611.47
##
## Top 3 models based on the ICL criterion:
##       V,2       E,2       V,3
## -28268.94 -28280.50 -28292.16
```

Rappresentiamo i valori ottenuti per l'ICL all'aumentare del numero di componenti per i due modelli. Anche in questo caso, come per il BIC, selezioneremo il numero di componenti corrispondente al valore più elevato dell'ICL.

```
icl.df = data.frame(G = rep(1:9,2),
                    ICL = c(icl[,"E"],icl[,"V"]),
                    modelNames = c(rep("E",9),
                                   rep("V",9)))

ggplot(icl.df,aes(x = G,y = ICL,col = modelNames)) +
  geom_line(aes(group = modelNames),size = 1) +
  geom_point(aes(shape = modelNames),size = 3) +
  geom_vline(xintercept = 2,col = 2,lty = 2,size = 1) +
  scale_shape_manual(values = c(19,17),
                     labels = c("E","V")) +
  scale_color_manual(values = paletteer_c("ggthemes::Classic Blue",6)[c(2,4)]) +
  scale_x_continuous(breaks = 1:9) +
  labs(x = "Numero di componenti",y = "ICL",shape = "",col = "") +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 10),
        legend.text = element_text(size = 10),
        legend.key.size = unit(1.5,"cm"))
```



Vengono selezionate $G = 2$ componenti e una struttura di varianze diverse tra i *cluster*. Possiamo notare come fino ad un numero di componenti pari a 2 si ottengono valori di ICL molto simili per entrambi i modelli. A partire da 3 componenti, invece, il modello `V` riporta valori nettamente superiori rispetto al modello `E`.

Stimiamo allora un modello di mistura gaussiano `V` con 2 componenti.

```
set.seed(123)
mbc2 = Mclust(data$charges,G = 2,modelName = "V")
summary(mbc2)
```

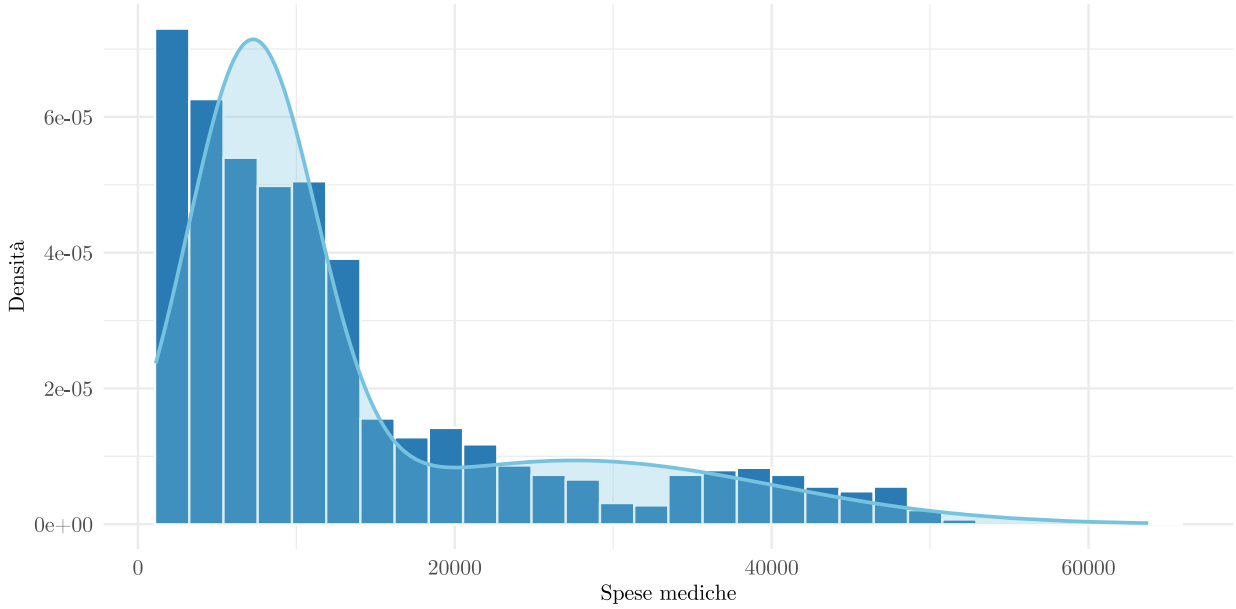
```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust V (univariate, unequal variance) model with 2 components:
##
##  log-likelihood    n df       BIC       ICL
##        -14018.6 1338  5 -28073.19 -28268.94
##
## Clustering table:
##   1   2
## 992 346
```

Il primo gruppo presenta una numerosità maggiore rispetto al secondo.

Confrontiamo la distribuzione osservata di `charges` con quella stimata dalla mistura di distribuzioni con 2 componenti.

```
df2 = data.frame(charges = data$charges,
                 density = mbc2[["parameters"]][["pro"]][1] *
                   dnorm(data$charges,mean = mbc2[["parameters"]][["mean"]][["1"]],
                         sd = sqrt(mbc2[["parameters"]][["variance"]]$sigmasq[1])) +
                   mbc2[["parameters"]][["pro"]][2] *
                   dnorm(data$charges,mean = mbc2[["parameters"]][["mean"]][["2"]],
                         sd = sqrt(mbc2[["parameters"]][["variance"]]$sigmasq[2])))

ggplot(df2,aes(x = charges,y = density)) +
  geom_histogram(aes(y = ..density..),col = "white",bins = 30,
                 fill = paletteer_c("ggthemes::Classic Blue",6)[4]) +
  geom_line(col = paletteer_c("ggthemes::Classic Blue",6)[2],size = 0.8) +
  geom_area(fill = paletteer_c("ggthemes::Classic Blue",6)[2],alpha = 0.3) +
  labs(x = "Spese mediche",y = "Densità") +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 10))
```



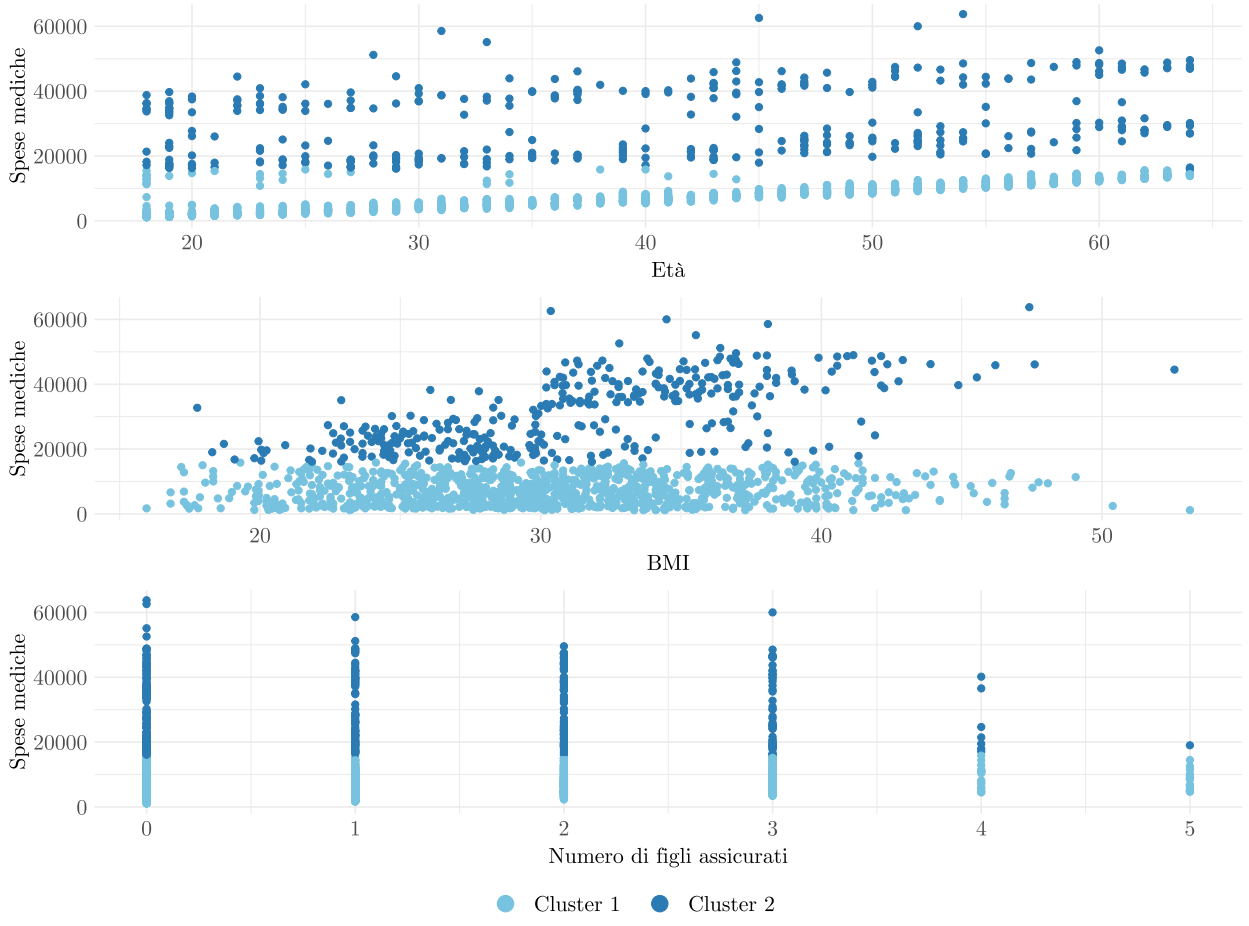
È evidente che l'adattamento della mistura alla distribuzione osservata sia peggiorato rispetto al modello con 5 componenti. Ora la distribuzione stimata anche nella coda sinistra, che prima presentava un ottimo adattamento alla distribuzione osservata, risulta meno precisa.

Rappresentiamo anche come le unità sono state suddivise nei due *cluster* considerando le variabili `age`, `bmi` e `children`.

```
g1 = ggplot(data,aes(x = age,y = charges,col = factor(mbc2[["classification"]]))) +
  geom_point(size = 2) +
  labs(x = "Età",y = "Spese mediche") +
  scale_color_manual(values = paletteer_c("ggthemes::Classic Blue",6)[c(2,4)],
                     label = paste("Cluster ",seq(1,2,1),sep = "")) +
  labs(col = " ") +
  guides(colour = guide_legend(override.aes = list(size = 5))) +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 15),
        axis.title = element_text(size = 15),
        legend.text = element_text(size = 15),
        legend.title = element_text(size = 15),
        legend.key.size = unit(1,"cm"))
g2 = ggplot(data,aes(x = bmi,y = charges,col = factor(mbc2[["classification"]]))) +
  geom_point(size = 2) +
  labs(x = "BMI",y = "Spese mediche") +
  scale_color_manual(values = paletteer_c("ggthemes::Classic Blue",6)[c(2,4)],
                     label = paste("Cluster ",seq(1,2,1),sep = "")) +
  labs(col = " ") +
  guides(colour = guide_legend(override.aes = list(size = 5))) +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 15),
        axis.title = element_text(size = 15),
        legend.text = element_text(size = 15),
        legend.title = element_text(size = 15),
        legend.key.size = unit(1,"cm"))
```

```r
g3 = ggplot(data,aes(x = children,y = charges,col = factor(mbc2[["classification"]]))) +
  geom_point(size = 2) +
  labs(x = "Numero di figli assicurati",y = "Spese mediche") +
  scale_color_manual(values = paletteer_c("ggthemes::Classic Blue",6)[c(2,4)],
                     label = paste("Cluster ",seq(1,2,1),sep = "")) +
  labs(col = " ") +
  guides(colour = guide_legend(override.aes = list(size = 5))) +
  theme_minimal() +
  theme(text = element_text(family = "CMUSerif"),
        axis.text = element_text(size = 15),
        axis.title = element_text(size = 15),
        legend.text = element_text(size = 15),
        legend.title = element_text(size = 15),
        legend.key.size = unit(1,"cm"))

ggarrange(g1,g2,g3,nrow = 3,common.legend = T,legend = "bottom")
```



Possiamo notare che il primo *cluster* con numerosità maggiore include individui per cui le spese mediche assumono un valore medio-basso mentre il secondo quelle con un valore più elevato di `charges`. Possiamo già anticipare che non sembra esserci una chiara suddivisione per età, BMI o numero di figli assicurati.

Valutiamo però, anche in questo caso, quali variabili hanno un effetto significativo su `charges` nei due *cluster* individuati.

```
lm1.2 = lm(charges ~ .,data = data %>%
             filter(mbc2[["classification"]] == 1))
summary(lm1.2)
```

```
##
## Call:
## lm(formula = charges ~ ., data = data %>% filter(mbc2[["classification"]] ==
##     1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1969.1  -602.9  -262.7   277.4 12799.9
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2961.539    247.975 -11.943  < 2e-16 ***
## age               254.975      3.017  84.505  < 2e-16 ***
## sexmale          -454.302     83.257  -5.457 6.14e-08 ***
## bmi                 9.511      7.241   1.313   0.1893
## children          421.241     34.263  12.294  < 2e-16 ***
## smokeryes       11750.920    390.203  30.115  < 2e-16 ***
## regionnorthwest  -333.659    119.131  -2.801   0.0052 **
## regionsoutheast  -658.951    123.931  -5.317 1.31e-07 ***
## regionsouthwest  -626.740    118.737  -5.278 1.60e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1310 on 983 degrees of freedom
## Multiple R-squared:  0.8911, Adjusted R-squared:  0.8902
## F-statistic:  1006 on 8 and 983 DF,  p-value: < 2.2e-16
```

```
lm2.2 = lm(charges ~ .,data = data %>%
             filter(mbc2[["classification"]] == 2))
summary(lm2.2)
```

```
##
## Call:
## lm(formula = charges ~ ., data = data %>% filter(mbc2[["classification"]] ==
##     2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20562.9  -5055.2    209.4   4648.5  30162.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -22010.68    2430.46  -9.056  < 2e-16 ***
## age                217.92      26.11   8.348 1.82e-15 ***
## sexmale             13.22     739.47   0.018    0.986
## bmi               1171.60      63.79  18.366  < 2e-16 ***
## children           -37.71     312.40  -0.121    0.904
## smokeryes         9977.52     858.75  11.619  < 2e-16 ***
```

```
## regionnorthwest    413.90    1062.05   0.390    0.697
## regionsoutheast   -926.02     981.63  -0.943    0.346
## regionsouthwest    431.38    1099.60   0.392    0.695
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6731 on 337 degrees of freedom
## Multiple R-squared:  0.6166, Adjusted R-squared:  0.6075
## F-statistic: 67.76 on 8 and 337 DF,  p-value: < 2.2e-16
```

Per il primo gruppo solamente la variabile `bmi` non risulta significativa, mentre per il secondo non lo sono `bmi`, `children` e `region`.

Salviamo l'appertenenza delle unità nei gruppi e consideriamo le variabili raggruppate in classi per valutare le caratteristiche dei *cluster*.

```
data2$latent_class2 = mbc2[["classification"]]
table(data2$age,data2$latent_class2)
```

```
##
##           1   2
##   18-24 213  65
##   25-34 201  70
##   35-49 297 107
##   50-64 281 104
```

```
table(data2$bmi,data2$latent_class2)
```

```
##
##                        1   2
##   Sottopeso           18   2
##   Normopeso          169  53
##   Sovrappeso         283  94
##   Obeso              299 100
##   Estremamente obeso 223  97
```

```
table(data2$children,data2$latent_class2)
```

```
##
##        1   2
##   0  439 135
##   1  243  81
##   2  167  73
##   3+ 143  57
```

```
table(data2$smoker,data2$latent_class2)
```

```
##
##         1   2
##   no  980  84
##   yes  12 262
```

```r
table(data2$region,data2$latent_class2)
```

```
##
##               1   2
##   northeast 239  85
##   northwest 248  77
##   southeast 249 115
##   southwest 256  69
```

```r
table(data2$sex,data2$latent_class2)
```

```
##
##             1   2
##   female 509 153
##   male   483 193
```

In generale, i due gruppi non evidenziano particolari caratteristiche per le variabili considerate. L'eccezione è data dalla variabile `smoker` per la quale possiamo notare che il primo gruppo include principalmente individui non fumatori, mentre il secondo gruppo include una maggioranza di fumatori.

Alla luce di queste considerazioni, possiamo affermare che il modello con 5 componenti risulta essere più adatto per descrivere la distribuzione delle spese mediche rispetto al modello con 2 componenti.

Si specifica che sono stati stimati anche dei modelli di mistura con 3 e 4 componenti, di cui non viene riportato il codice. La suddivisione a 3 gruppi risulta simile a quanto ottenuto con il modello selezionato con l'ICL a 2 componenti anche se si ottiene una suddivisione più ragionevole degli assicurati per classi di età. Al contrario, un modello di mistura con 4 componenti riporta una suddivisione paragonabile a quella ottenuta con il modello a 5 componenti.

Conclusioni

In questa analisi, è stato preso in considerazione un insieme di dati relativo ad individui americani iscritti ad un piano assicurativo e contenente, oltre ad alcune loro caratteristiche, le spese mediche addebitate a compagnie assicurative.

È stata utilizzata la metodologia del *model-based clustering* al fine di individuare possibili sottopopolazioni di assicurati in base al valore delle spese mediche. Il modello selezionato dal criterio BIC presenta 5 componenti e prevede una struttura di varianze diverse tra i *cluster*. Questo modello ha mostrato un buon adattamento alla distribuzione osservata delle spese mediche, in particolare per i valori più elevati. Successivamente è stata valutata la significatività delle variabili all'interno dei gruppi individuati. In generale, la variabile che sembra influenzare maggiormente il valore delle spese mediche è il fatto di essere fumatori. Considerando la variabile `smoker`, infatti, si osserva una suddivisione ragionevole degli individui nei gruppi individuati.

È stato poi considerato un modello di mistura con 2 componenti, selezionato tramite l'ICL. L'individuazione di due soli gruppi è stata, tuttavia, valutata troppo semplicistica e poco adatta a cogliere le sottopopolazioni presenti nei dati.

In conclusione, possiamo affermare che il modello di mistura Gaussiana con 5 componenti, selezionato tramite il BIC, risulta essere il più appropriato per analizzare il valore delle spese mediche addebitate alle compagnie assicurative, nonostante le sottopopolazioni non siano chiaramente distinte rispetto alle altre variabili. Tuttavia, l'analisi ha rivelato che i fumatori presentano un valore delle spese mediche significativamente più elevate, suggerendo di valutare un possibile incremento delle tariffe assicurative per questa categoria. Al contrario, i giovani non fumatori presentano costi inferiori, indicando che potrebbero beneficiare di una riduzione delle tariffe assicurative.

Bibliografia

[1] L. Scrucca, M. Fop, T. B. Murphy, et al. “mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models”. In: *The R Journal* 8 (1 2016). https://doi.org/10.32614/RJ-2016-021, pp. 289-317. ISSN: 2073-4859. DOI: 10.32614/RJ-2016-021.