Healthcare Smokers Analysis

Sveva Maria Martilotti

INTRODUCTION

Among the various factors that influence healthcare costs, smoking has long been recognized as one of the primary negative determinants. However, despite widespread awareness of the harmful effects of smoking, the differences in healthcare costs between smokers and non-smokers are not always clear, especially when the sample of smokers is relatively small compared to non-smokers.

In this project, we will analyze a sample of people from a population divided into two groups: **smokers** and **non-smokers**. The data used for this project are sourced from *Kaggles*.

Goals

We aim to:

- Calculate and compare the average health expenditure (FIRST PART)
- Examine the impact of BMI, number of children, and smoking status on healthcare costs within each region, seeking to understand how these factors interact with smoking status (SECOND PART)

The main challenge lies in the fact that the sample of smokers is significantly smaller compared to that of non-smokers (Table 1), which could affect the precision of the estimates. To address this challenge, we will use a *Bayesian hierarchical model*, which allows us to obtain more "robust" estimates: to reduce the uncertainty of the estimates, the model leverages data from other groups to derive the posterior means for each group (**shrinkage effect**: the estimates for individual groups tend to an overall mean).

This study will not only contribute to a better understanding of the differences in healthcare costs between smokers and non-smokers but also provide valuable insights into how variables such as BMI, number of children, and age influence these costs, supporting potential targeted health interventions and policies.

```
##   smoker   mean     sd n_observations
##   <chr>   <dbl>  <dbl>          <int>
##  no      8434.  5994.           1064
##  yes    32050. 11542.            274
```

Table 1: it shows the sample mean and the standard deviation and the sample size

Hierarchical normal model

Assumptions

We treat the data within the groups as being conditionally i.i.d. given a parameter, which we call **within-group sampling variability**:

$$\{Y_{1,j},\dots,Y_{n_j,j} \mid \phi_j\} \sim \text{i.i.d. } p\big(y \mid \phi_j\big)$$

Where $j = 1,\dots\dots,m.$ are the groups and $i = 1,\dots,n_j$ are the obs in each group.

Further, we have many groups with parameters ϕ_j that we assume are sampled from a population of *groups* we can treat the group means ϕ_j as conditionally i.i.d. given another parameter, which we call **between-group sampling variability**:

$$\{\phi_1,\dots,\phi_m \mid \psi\} \sim \text{i.i.d. } p(\phi \mid \psi)$$

The key of the hierarchical normal model is to treat the data within a group as being normally distributed with some mean θ_j and variance σ^2, and the means among groups to *also* be normally distributed according to some other mean μ and variance τ^2.

Then, we simply need a prior distribution on the parameter ψ :

$$\psi \sim p(\psi)$$

Note: in this project, we're assuming that the data within groups share a common variance σ^2 that doesn't depend on the group *j*

Model Specification (First Part)

We consider m independent groups, each one of them with nj independent normally distributed data points, $y_{i,j}$, each of which with group-specific mean $\mu_{i,j}$and common variance σ^2.

$$y_{1,j}.\dots.y_{nj,j} \mid \mu_j,\sigma^2 \overset{\text{ind}}{\sim} N\big(\boldsymbol{\mu_j},\sigma^2\big), \quad i = 1,\dots,n_j, \quad j = 1,2$$

In addition, we propose a hierarchical prior distribution with the following stages:

$$\boldsymbol{\mu_j} \mid \boldsymbol{\mu},\tau \sim N(\boldsymbol{\mu},\tau) \quad \text{and} \quad \sigma^2 \overset{\text{iid}}{\sim} \text{IG}(\alpha, b),$$

With $\boldsymbol{\mu} \sim N(\boldsymbol{\mu}_0, \Lambda_0)$ $\quad\tau \sim$ IG (α_ τ, b_τ)

Prior hyperparameters

$\mu_0 = 20.000$	$\Lambda_0 = 15.000$
α_τ = 0.01	b_τ = 0.01
α_ = 0.01	b = 0.01

We had set weakly informative and non informative priors.

MCMC approximation

In order to approximate to various posterior features of interest, specially the posterior mean, we have developed the following JAGS scrips:

```
# 1. Write the model as a string
modelString <- "
model {
  ## sampling
  for (i in 1:N) {
    y[i] ~ dnorm(mu_j[Index[i]], invsigma2)
  }
  ## priors
  for (j in 1:J) {
    mu_j[j] ~ dnorm(mu, invtau2)
  }
  invsigma2 ~ dgamma(a, b)
  sigma <- sqrt(1 / invsigma2)
  ## hyperpriors
  mu ~ dnorm(mu0, L0)
  invtau2 ~ dgamma(a_t, b_t)
  tau <- sqrt(1 / invtau2)
}
"
# 2. Prepare the data for the model
y <- df$charges
Index <- df$smoker
N <- length(y)
J <- length(unique(Index))

the_data <- list(
  "y" = y,
  "Index" = Index,
  "N" = N,
  "J" = J,
  "mu0" = 20000,
  "L0" = 15000,
  "a_t" = 0.01,
  "b_t" = 0.01,
  "a" = 0.01,
  "b" = 0.01
)
```

```
# 3. Create the model
jags_model <- jags.model(textConnection(modelString),
                         data = the_data,
                         n.chains = 1,
                         n.adapt = 5000)
# Burning of 50.000
update(jags_model, n.iter = 50000)

# Sampling from the posterior after the burning and with a thinning of 10
samples <- coda.samples(jags_model,
                        variable.names = c("mu", "tau", "mu_j", "sigma"),
                        n.iter = 250000,
                        thin=10)
```

Convergence diagnostic MCMC

The output of an MCMC chain is a dependent sequence.

Since the correlation in the sequence can affect the approximations and the ability of the chain to converge, It's good to check for convergence.

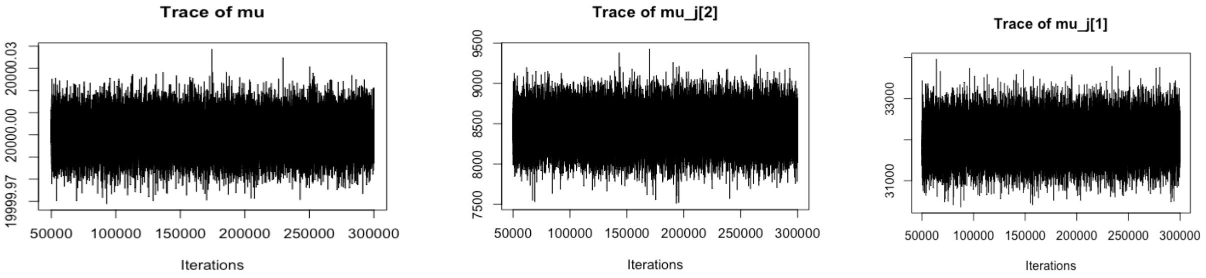


Figure 2: Trace plot

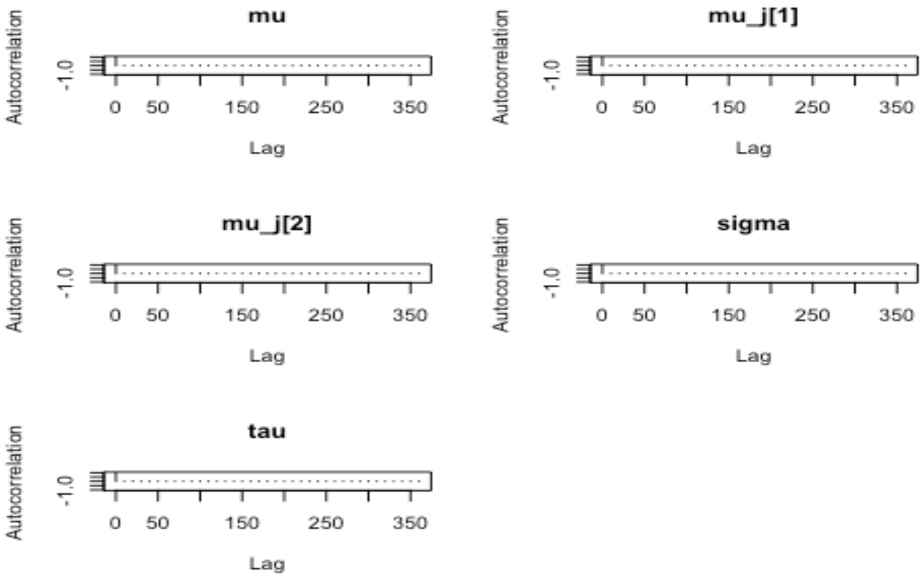


Figure 3: Autocorrelation plot. It measures the correlation in the sequence at different lags

From Figure 2, we can see that the values do not seem to follow a specific trend and that the values do not seem to be getting stuck in any particular region. Figure 3 indicates that the samples are almost independent.

In general, we can say that the chain has reached good convergence.

```
# Results:
summary(samples)
##
## Iterations = 50010:3e+05
## Thinning interval = 10
## Number of chains = 1
## Sample size per chain = 25000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
```

```
##          Mean        SD  Naive SE Time-series SE
## mu      20000 8.156e-03 5.158e-05      5.158e-05
## mu_j[1] 32032 4.527e+02 2.863e+00      2.863e+00
## mu_j[2]  8440 2.290e+02 1.449e+00      1.449e+00
## sigma    7475 1.459e+02 9.229e-01      9.229e-01
## tau     20467 3.153e+04 1.994e+02      2.042e+02
##
## 2. 95% CI for each variable:
##
##          2.5%   97.5%
## mu      20000   20000
## mu_j[1] 31137   32915
## mu_j[2]  7987   8885
## sigma    7195   7763
## tau      6125   71681
```

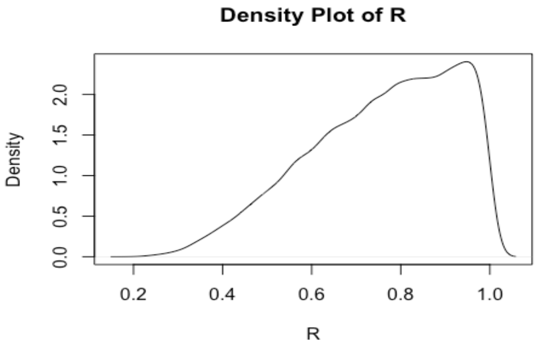
From the results just reported, we can say that:

- Smokers have an average healthcare expenditure of about 32.032 \$ (μ_1), while non-smokers have an average of abou 8440\$ (μ_2), confirming that smokers tend to have higher healthcare costs. The fact that their confidence intervals (CIs) do not overlap suggests a statistically significant difference between the average costs of the two groups.

- Sigma, the variability within each group, is quite high. This suggests that there are many factors besides smoking that influence individual healthcare costs

- Tau, the variability between groups, is very significant with a very wide confidence interval (high uncertainty in the estimate of tau).

To compare these two sources of variation, we have computed $R = \frac{\boldsymbol{\tau}^2}{\sigma^2+\boldsymbol{\tau}^2}$ che rappresenta la parte di variabilità totale nelle spese sanitarie dovuta alla differenze between groups.



```
R_cred_interval
##      2.5%     97.5%
## 0.4020247 0.9892307
```

Figure 4

From Figure 4, we can see that most of the posterior probability of R is concentrated on the right side, indicating that the variability in the data is mainly due to the variability between groups, confirming the significance of the hierarchical structure.

The considerable variability within the groups indicates that personalized strategies may be necessary to manage individual healthcare costs. It should also be noted that a **regional effect** (region of residence of individuals: *northeast*, *northwest*, *southeast*, and *southwest*) was considered, but it was found that the costs within each group are very similar across the regions, and therefore, we deemed it not to have a significant impact within the groups.

Based on this, in the following part of the analysis, it was decided to keep the data separate between smokers and non-smokers and to add the following variables to examine the significance of their impact on healthcare costs.

1. **Age:** The age of the patient, whose range of values is 18 – 64, hence all the population is condierares as adult.
2. **Bmi:** The Body Mass Index (BMI) of the patient. Since the age of the population, we have taked into consideration the following standards:

BMI	Weight Status
< 18.5	Underweight
18.5—24.9	Healthy Weight
18.5—24.9	Overweight
≥ 30	Obesity

3. **Children:** The number of children or dependents covered under the medical insurance, influencing family-related medical expenses.

Hierarchical normal linear regression model (Second Part)

Model Specification

We consider m independent groups, each one of them with nj independent normally distributed data points, $y_{i,j}$, each of which with group-specific mean $\mu_{i,j} = \mathbf{x}_{i,j}^{\top}\boldsymbol{\beta}_j$, with $\boldsymbol{\beta}_j = \left(\beta_{1,j}, \dots, \beta_{p,j}\right)$, and common variance σ^2; i.e.,

$$y_{1,j}.\dots.y_{nj,j} \mid \boldsymbol{\beta}_j, \sigma^2 \overset{\text{ind}}{\sim} N\left(\mathbf{x}_{i,j}^{\top}\boldsymbol{\beta}_j, \sigma^2\right), \quad i = 1, \dots, n_j, \quad j = 1, \dots, m.$$

In addition, we propose a hierarchical prior distribution with the following stages:

$$\boldsymbol{\beta}_1, \dots.., \boldsymbol{\beta}_m \mid \boldsymbol{\theta}, \Sigma \overset{\text{iid}}{\sim} N_p(\boldsymbol{\theta}, \Sigma) \quad \text{and} \quad \sigma^2 \overset{\text{iid}}{\sim} \text{IG}(\nu_0/2, \nu_0\xi_0^2/2),$$

With $\boldsymbol{\theta} \sim N_p(\boldsymbol{\mu}_0, \Lambda_0) \quad \Sigma \sim IW_p(\eta_0, \mathbf{S}_0^{-1})$

where IW denotes the Inverse Wishart distribution and G denotes the Gamma distribution.

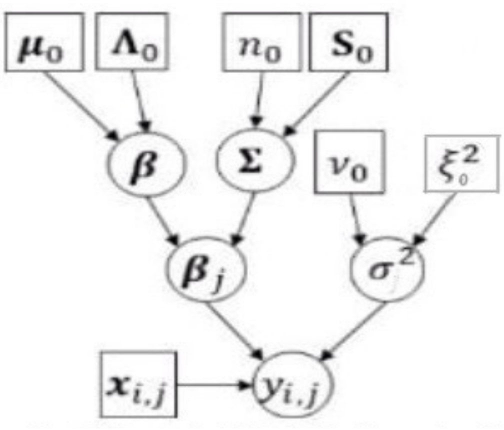


Figure 5: DAG representation of the hierarchical Normal lineare regression model

1: Prior hyperparameters

In order to compute the hyperparameters values, we have used a ***unit information prior*** (UIP) approach, considering it a good compromise between a non-informative and informative prior, being able to influence the estimates without dominating too much the data evidence.

They have been obtained as follows:

μ_0 = average of $\hat{\beta}_{OLS}$. It's a vector of p element.

Λ_0= symmetric matrix p x p of the sample covariance of $\hat{\beta}_{OLS}$.

ν_0= 1

ξ_0^2 = average of the within - group sample variance $\widehat{\sigma 2}_{OLS}$

η_0= p+2

S_0= Λ_0

Note: p is the number of predictors

2: Posterior Inference

Joint posterior inference for the model parameters can be achieved by a Gibbs sampling algorithm, which requires iteratively sampling each parameter from its full conditional distribution.

Let $\Theta = (\sigma^2, \beta_j, \theta, \Sigma)$ the full set of paramters in the model. The posterior distribution of Θ is

$$p(\Theta|y_{i,j}) \propto p(y_{i,j}|\beta_j, \sigma^2)p(\beta_j|\theta, \Sigma)p(\theta)p(\Sigma)p(\sigma^2)$$

which leads to

$$p(\Theta|y_{i,j}) \propto$$

$$\prod_{j=1}^{m} \prod_{i=1}^{n_j} \sigma^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\left(y_{i,j} - x_{i,j}^{\top}\beta_j\right)^2\right)$$

$$\times \prod_{j=1}^{m} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}\left(\beta_j - \theta\right)^{\top} \Sigma^{-1}\left(\beta_j - \theta\right)\right)$$

$$\times \exp\left(-\frac{1}{2}(\theta - \mu_0)^{\top} \Lambda_0^{-1}(\theta - \mu_0)\right)$$

$$\times |\Sigma|^{-(\eta_0+p+1)/2} \exp\left(-\frac{1}{2} tr(S_0 \Sigma^{-1})\right)$$

$$\times (\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)} \exp\left(-\frac{\nu_0 \xi_0^2}{2\sigma^2}\right)$$

Thus, we have that:

The full conditional distribution of $\boldsymbol{\beta_j}$, with j=1,…,m, is

$$\beta_j|\text{rest}\sim N_p\Big((\Sigma^{-1}+\sigma^{-2}X^{\top}{}_j X_j)^{-1}(\Sigma^{-1}\theta+\sigma^{-2}X^{\top}{}_j y_j) , (\Sigma^{-1}+\sigma^{-2}X^{\top}{}_jX_j)^{-1}\Big)$$

The fcd of $\boldsymbol{\theta}$ is

$$\theta|\text{rest}\sim N_p\Big((\Lambda^{-1}{}_0+m\,\Sigma^{-1})^{-1}\,(\Lambda^{-1}{}_0\,\mu_0+\Sigma^{-1}\textstyle\sum_{j=1}^{m}\beta\text{j}) , (\Lambda^{-1}{}_0+m\,\Sigma^{-1})^{-1}\Big)$$

The fcd of Σ is

$$\Sigma \mid \text{rest}\sim IW\ \Big(\eta_0+m\, ,\ (S_0+\sum_{j=1}^{m}(\beta\text{j}-\theta)(\beta\text{j}-\theta)^{\top})^{-1}\Big)$$

The fcd of $\boldsymbol{\sigma^2}$ is

$$\sigma^2 \mid \text{rest}\sim IG\ \Big(\frac{v_0+\sum \text{nj}}{2} , \frac{v_0\ \xi^2{}_0+\sum_j\sum_i(y_{ij}-x^{T}{}_{ij}\beta_j)^2}{2}\ \Big)$$

Let's $\phi^{(s)}$ denote the state of parameter ϕ in the s-th iteration of the Gibbs sampling algorithm, for s = 1,…,S. Then, the algorithm works as follows:

1. Choose an initial value for each parameter in the model, say $\beta^{(0)}{}_1,\ldots,\beta^{(0)}{}_m$, $\theta^{(0)}$, $\Sigma^{(0)}$,$\sigma^{2(0)}$
2. For s= 1,…,S update each parameter:

   a. Sample $\beta^{(s)}{}_j$ from its fcd $p(\ \beta_j \mid \theta^{(s-1)}\ , \Sigma^{(s-1)}\ ,\sigma^{2(s-1)}\ ,y_j)$

   b. Sample $\theta^{(s)}$ from its fcd $p(\ \theta \mid \beta^{(s)}{}_j,\ \Sigma^{(s-1)})$

   c. Sample $\Sigma^{(s)}$ from its fcd $p(\ \Sigma \mid \beta^{(s)}{}_j\ , \theta^{(s)})$

   d. Sample $\sigma^{2(s)}$ from its fcd $p\ (\ \sigma^2 \mid \beta^{(s)}{}_j,y_{ij})$

Explorative analysis

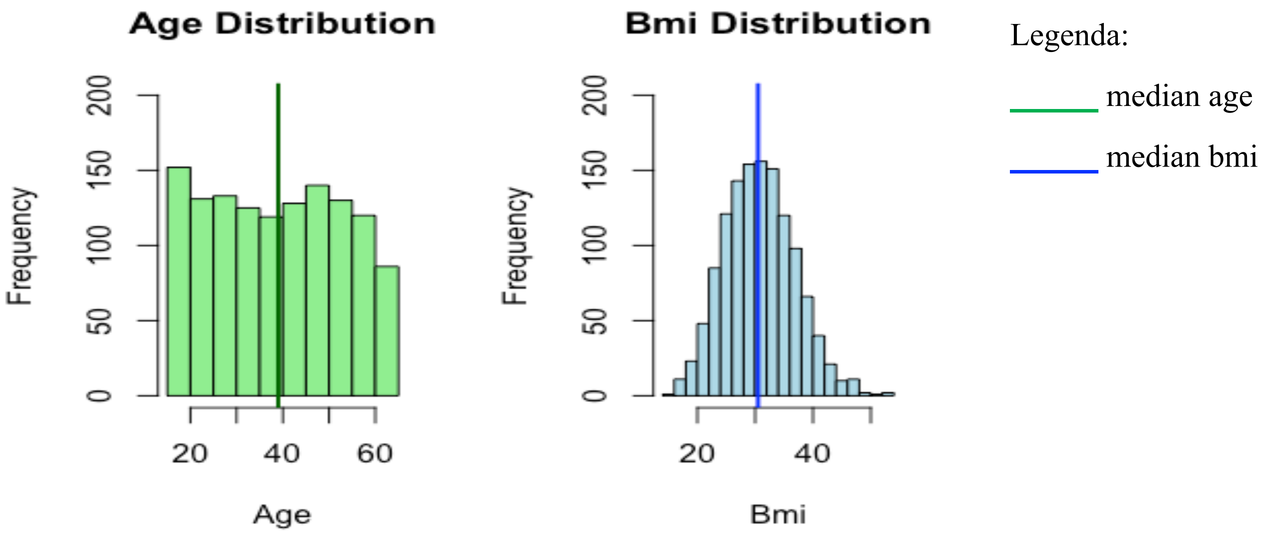


Figure 6

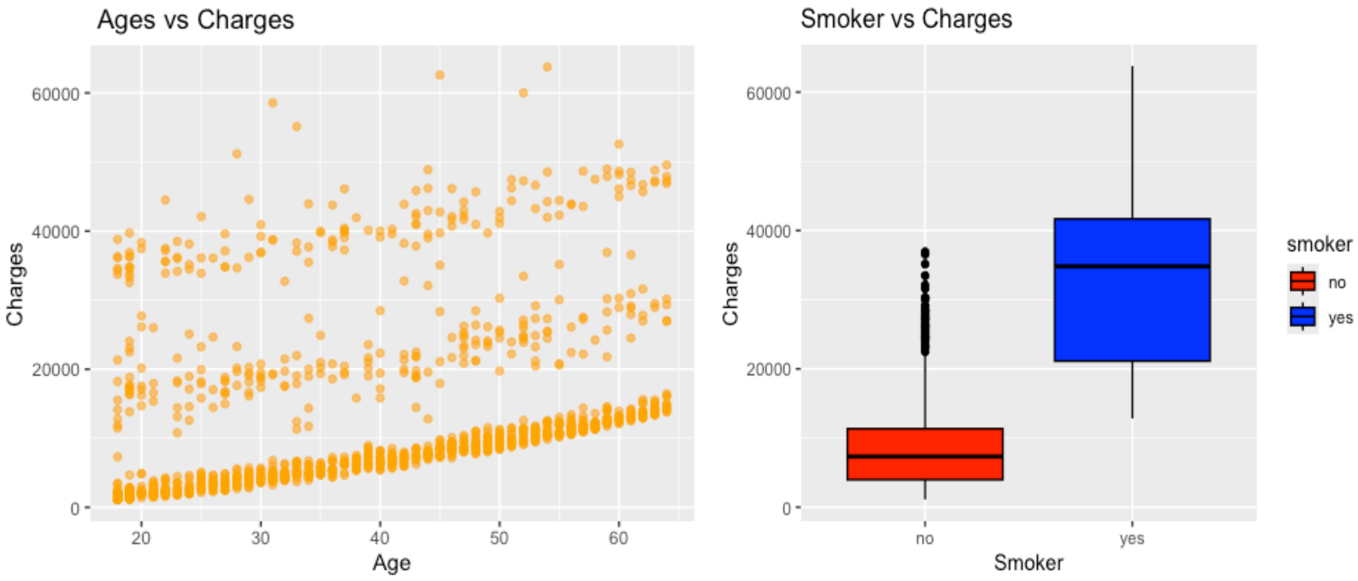


Figure 7

According to this initial analysis, we know that:

- The age distribution is fairly symmetrical, with a slight right skew, meaning that the number of young people in the population is slightly higher than that of the elderly (from Figure 6).
- The BMI distribution shows that most people have a BMI between 25 and 35. This is consistent with the well-known fact that the obesity rate dominates in the USA (from Figure 6).
- The scatterplot shows a positive correlation between age and expenses. This means that, in general, older people have higher expenses than younger people (from Figure 7).

```
# Empty list
X <- list()
Y <- list()
fit <- list()
S2_LS <- numeric()

# N. groups
m <- 2

for (j in 1:m) {
  # Extract data for group j
  data_region <- df[df$smoker == j, ]

  # Extract predictors and the response variabile
  predictors <- data_region[, c("bmi", "children","age")]
  predictors <- scale(predictors, center = TRUE, scale = TRUE)
  predictors_df<- as.data.frame(predictors)
  response<- scale(data_region$charges, center = FALSE, scale = TRUE)
  Y[[j]] <- response

  # Design matrix
  X[[j]] <- model.matrix(~ bmi + children + age, data = predictors_df)

  # fit a linar regression model for each group j
  fit[[j]] <- lm(Y[[j]] ~ -1 + X[[j]])
  # Calcola le stime dei residui e la varianza degli errori
  S2_LS <- c(S2_LS, summary(fit[[j]])$sigma^2)
}

BETA_LS <- do.call(rbind, lapply(fit, function(model) {
  coef(model)
}))
```

Note: BETA_LS is a matrix m x p, hence for each group it shows the OLS estimated coefficients.

MCMC

```
# preparing the values
theta<-mu0<-apply(BETA_LS,2,mean)
s2<-s20<-mean(S2_LS)
L0<-as.matrix(cov(BETA_LS))
eta0<-p+2
epsilon <- 1e-5
L0 <- L0 + diag(epsilon, nrow(L0))
Sigma<-S0<-L0
BETA<-BETA_LS
THETA.b<-S2.b<-NULL
iL0<-solve(L0)
iSigma<-solve(Sigma)
Sigma.ps<-matrix(0,p,p)
SIGMA.PS<-NULL
BETA.ps<-BETA*0
BETA.pp<-NULL
nu0=1
```

```r
set.seed(1)
for (s in 1:25000){
  #update beta_j
  for(j in 1:m){
    Vj<-solve(iSigma+t(X[[j]])%*%X[[j]]/s2)
    Ej<-Vj%*%(iSigma%*%theta+t(X[[j]])%*%Y[[j]]/s2)
    BETA[j,]<-rmvnorm(1,Ej,Vj)
  }
  #update theta
  Lm<-solve(iL0+m*iSigma)
  mum<-Lm%*%(iL0%*%mu0+iSigma%*%apply(BETA,2,sum))
  theta<-t(rmvnorm(1,mum,Lm))

  #update Sigma
  mtheta<-matrix(theta,m,p,byrow=TRUE)
  iSigma<-rwish(1,eta0+m,solve(S0 + t(BETA - mtheta) %*% (BETA - mtheta)))


  #update s2
  RSS<-0
  for(j in 1:m){
    RSS<-RSS+sum((Y[[j]]-X[[j]]%*%BETA[j,])^2)
  }
  s2<-1/rgamma(1,(nu0+1338)/2,(nu0*s20+RSS)/2)

  # store result with thinning of 10
  if(s%%10==0){
    cat(s,s2,"\n")
    S2.b<-c(S2.b,s2)
    THETA.b<-rbind(THETA.b, t(theta))
    Sigma.ps<-Sigma.ps+solve(iSigma)
    BETA.ps<-BETA.ps+BETA
    SIGMA.PS<-rbind(SIGMA.PS,c(solve(iSigma)))
    BETA.pp<-rbind(BETA.pp,rmvnorm(1,theta,solve(iSigma))) #post pred distr
  }
}
```

Finally, we have computed all the values we needed to initialize the chain and now we run a Gibbs sampler for 25.000 draws and saves every 10th draws, obtaining a sequence of 2.500 values for each parameter.

Output:

- **THETA.b** stores the updated samples of θ during each iteration that is recorded. It has a dimension of *2500 x p*.

- **SIGMA.PS** stores the updated samples of Σ during each iteration that is recorded. It has a dimension of *2500 x (p x p)*, where p x p is the number of elements in a covariance matrix.

- **S2.b** stores the updated samples of ξ_0^2 during each iteration that is recorded. It is a sequence of *2500* values.

- **BETA.ps** stores the cumulative sum of BETA for each iteration that is recorded, where BETA is a matrix *m x p* representing the estimated coefficients for each person j and for each predictor.

At each iteration, BETA is updated and the sum of the BETA values at each iteration is accumulated in BETA.ps.

We can use the simulated values to make Monte Carlo approximation to various posterior features of interest.

Convergence diagnostic MCMC

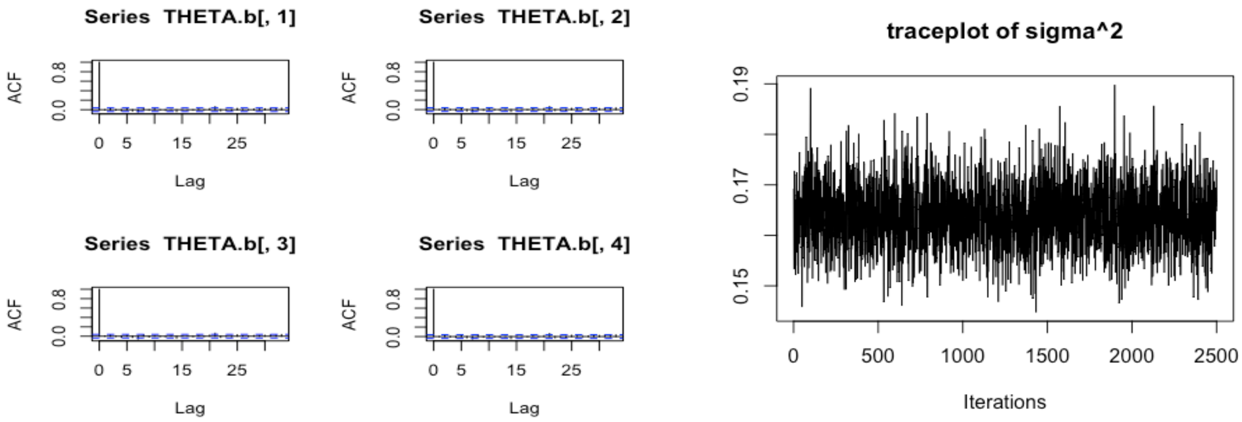


Figure 8: Autocorrelation at different lags and treceplot

```
# Print Geweke test results
print(geweke_theta)
##
## Fraction in 1st window = 0.1
## Fraction in 2nd window = 0.5
##
## X[[j]](Intercept)         X[[j]]bmi    X[[j]]children          X[[j]]age
##             1.727             1.819            -1.923             -1.702
print(geweke_s2)
##
## Fraction in 1st window = 0.1
## Fraction in 2nd window = 0.5
##
##  var1
## 1.011
```

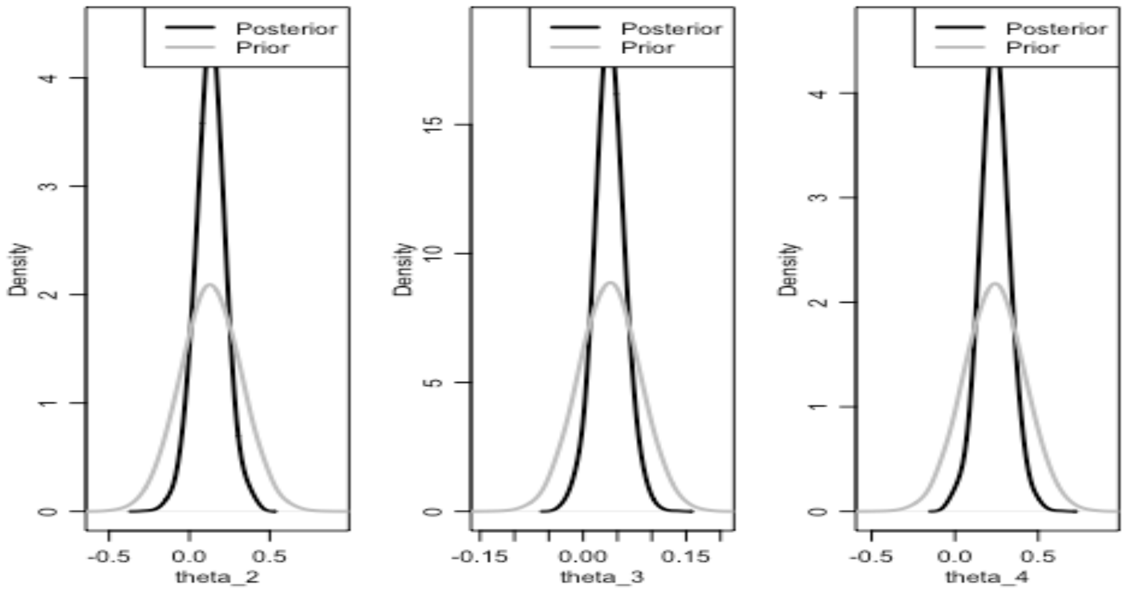
The Geweke test compares the mean of the 1st window with that of the 2nd window to determine if the MCMC chain is converging.
The idea is that if the chain has reached convergence, then if I take a statistic calculated at two opposite area of the chain, these will not be significantly different.

In general, the traceplot, the autocorrelation plot and the Geweke test confirm that the chain is converging correctly.

Posterior means and CI for the regressor parameters in the hierarchical Normal linear regression model

Below, we can see the posterior mean, the standard deviation (uncertainty of the estimate), and the credible intervals (CI), which represent the range within the true value of the parameter lies with a certain probability (95%), for each coefficient.

```
##                 Mean         SD         2.5%      97.5%
## Intercept 0.87549567 0.04375198  0.786707523 0.96563798
## Bmi       0.13092589 0.09220241 -0.055310786 0.32460919
## children  0.03850063 0.02164478 -0.006059069 0.08292162
## age       0.23826749 0.08869575  0.052579355 0.41573578
```



The **prior distribution** represents the initial knowledge of the parameter before observing the data: it is more spread since the uncertainty about the parameter is greater.

The **posterior distribution** reflects our updated knowledge of the parameter based on the observed data: it is narrower and taller compared to the prior distribution, indicating that the information from the data has reduced the uncertainty, and it is centered around the estimated value of the parameter.

What about the variance?

```
## [1] "posterior mean of sigma^2:"
## [1] 0.1639662

## [1] "CI for sigma^2:"
##      2.5%     97.5%
## 0.1518682  0.1764108
```

Model checking

We control whether the model is good enough for the purposes of the analysis. We check the quality of the model by drawing simulated values from the posterior predictive distribution of replicated data and comparing these samples to the observed data.

Specifically, we have calculated the **Posterior Predictive p-value** (***PPP-value***), a measure used to evaluate the quality of the model by comparing the observed data with the data replicated by the model, given the posterior parameters. In particular, *it's the probability that the replicated data could be more extreme than the observed data.*

$$ppp = \Pr(t(y^{rep}) > t(y) | \, y)$$

Where y^{rep} is a predictive dataset and t a test statistic (e.g. the mean).

A PPP-value of 0.502 is very close to 0.5, which is the reference value for a uniform distribution. This suggests that there is no strong evidence of model misfit.

CONCLUSION

In this project, we employed a Bayesian hierarchical model to analyze healthcare costs among smokers and non-smokers, focusing on the impact of BMI, age, and the number of children on these costs. The posterior means of the regressors provided valuable insights into how these factors influence healthcare expenses: the results indicated that **smokers** have significantly higher healthcare costs, with a posterior mean of abour $32,032 compared to $8,440 for **non-smokers**.

The posterior means for the predictors indicated that **age** has the strongest positive association with healthcare costs, with a posterior mean of 0.238, and its credible interval doesn't include zero, confirming its significant impact. Given the inclusion of zero in the CI for **BMI** and the number of **children**, it's essential to interpret these results with caution. These predictors might not have a significant impact on healthcare costs based on the current model. Therefore, the effects of BMI and the number of children may require further investigation with more data or alternative models to clarify their roles.

The model showed a relatively low **posterior variance** ($\sigma^2 = 0.164$), suggesting that the healthcare costs are fairly predictable based on the included factors.

According to the ppp – value the model fits the data well.

The significant difference in healthcare costs between smokers and non-smokers highlights the potential benefits of smoking cessation programs. Additionally, the positive association between age and costs suggests that age-specific health interventions could be valuable in managing rising healthcare expenses.