# Risk factors for lung cancer per country in 2015

Aina Belloni & Francesca Ghidini

## Contents

# 1 Dataset

The dataset comes from the www.who.int

The data contains general information and some factors of health conditions for 183 countries, in particular there are 15 variables, which are:

- **Region**: 6 levels factor: Africa, Americas, Eastern Mediterranean, Europe, South-East Asia, Western Pacific
- **Death_ratio**: Ratio between the deaths caused by trachea, bronchus and lung cancers and the population in 2015
- **Males**: Percent of males citizens
- **Income**: World Bank income group. 4 levels factor that can be: High-income, Low-income, Lower-middle-income and Upper-middle-income depending on the value of the gross national income (GNI) per capita valued annually in US dollars using a three-year average exchange rate. The cutoff points between each of the groups are fixed in real terms: they are adjusted each year in line with price inflation. The classification is published on www.data.worldbank.org.
- **Government_expenditure**: 2015 health expenditure per capita in US$
- **Domestic_expenditure**: 2015 domestic private health expenditure per capita in PPP int$
- **Sanitations**: Percentage of population using at least basic sanitation services, that is, improved sanitation facilities that are not shared with other households.
- **Tobacco**: Estimate of current cigarette smoking prevalence (%) in 2015 (age-standardized rate[1]). It is the percentage of the population aged 15 years and over who currently use any tobacco product (smoked and/or smokeless tobacco) on a daily or non-daily basis. This is an estimate obtained with a statistical model based on a Bayesian negative binomial meta-regression. Further information on the method used are available in a peer-reviewed article in The Lancet, volume 385, No. 9972, p966–976 (2015). This variable presents Not Available data.
- **Age_Mean**: Population median age (years)
- **Alcohol**: Alcohol, recorded per capita (15+ years) consumption (in litres of pure alcohol)
- **BMI**: Mean body mass index (BMI) in kg/m2 of defined population (age-standardized estimate)
- **Pollution**: Annual mean concentration of particulate matter of less than 2.5 microns of diameter (PM2.5) [ug/m3] in urban areas
- **UV**: Average daily ambient ultraviolet radiation (UVR) level (in J/m2)
- **Water**: Population using at least basic drinking-water services (%) in 2015, that is, the population that drinks water from an improved source, provided collection time is not more than 30 minutes for a round trip.
- **SCI**: Coverage of essential health services (defined as the average coverage of essential services based on tracer interventions that include reproductive, maternal, newborn and child health, infectious diseases, non-communicable diseases and service capacity and access, among the general and the most disadvantaged population). The indicator is an index reported on a unit-less scale of 0 to 100

## 1.1 Goals

The main goal of this analysis is to explore possible lung cancer risk factors using our prior knowledge. To do that we rely on 2015 data that summarize peculiar features of a country such as sanitary system, wealth, demographic aspects and people habits.

Hence, the dependent variable in this case is the ratio between the number of deaths caused by lung cancer deaths and the total population, for each country.

## 1.2 Exploratory analysis

The variable of interest is represented in Figure 1. As it is possible to observe from Figure 1, the response variable has really small positive values, between 0 and 1, and its density distribution is bimodal and not symmetric.

[1] "Age-standardized rates account for the differences in the age structure of the populations being compared. In the calculation of the age-standardized rate, either one population is mathematically adjusted to have the same age structure as the other. In this way, the two groups are given the same age distribution structure so that a more representative picture of the characteristic in question is provided." (www.statcan.gc)

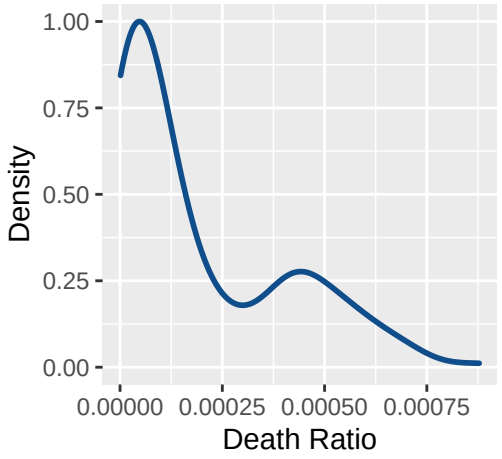


Figure 1: Density of the response variable Death ratio

### 1.3 Categorical variables

By doing more in-depth analysis we have noticed that there is a substantial difference between the mortality rate between the developing countries and the advanced countries. In fact, the bimodality of the response could be caused by the richest countries, for example those in Europe and those with a high income level, as we can notice in Figure 2. One probable reason could be the fact that the healthcare systems of poor regions is really underdeveloped compared to the ones of Europe and therefore the data could be less reliable and accurate.

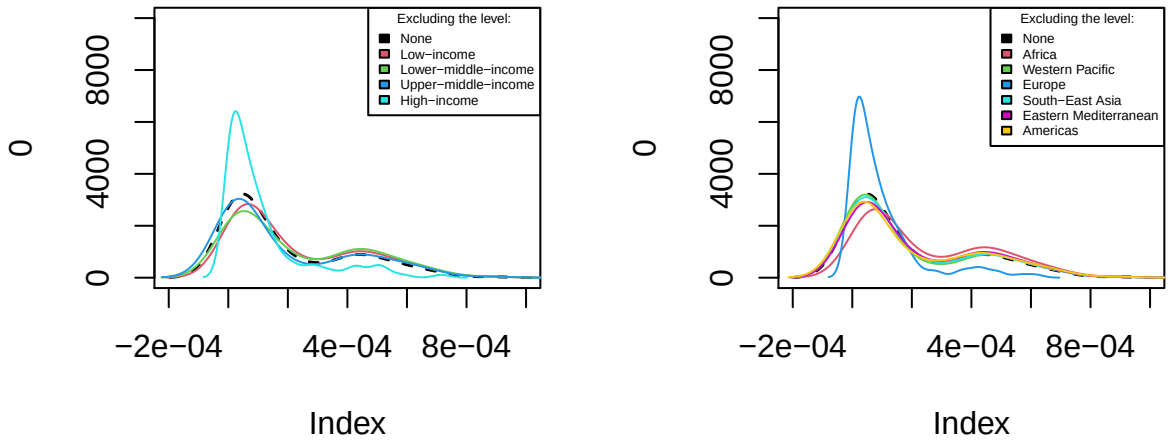


Figure 2: On the left: Density comparison of the response variables excluding different levels of the variable Income. On the right: Density comparison of the response variables excluding different levels of the variable Region

Since we have discovered that the possible cause of bimodality is given by the Region and the level of Income of each country we decided to explore in more depth the differences between these different groups.

First of all, we have calculated mean and variance for each level of `Income` and for each level of `Regions` and the total values as reported in the Table 1.

Figure 3 illustrates that between the various groups we get quite distinct values. A substantial difference is noted in particular with respect to the Africa and Low-income groups that are really different compared to the others level of the same factor. However, to be sure that our insight is correct we decided to test it with `Anova` using the F-test. The Anova consists in analyzing the variance; in particular, what we are interested in is the variation within and between groups that we call respectively:

Table 1: Summary Region and Income

| Level | Mean | Var | N |
|---|---|---|---|
| Africa | 2.418723e-05 | 1.3290e-09 | 47 |
| Europe | 4.382120e-04 | 3.2606e-08 | 50 |
| Americas | 1.418061e-04 | 1.9975e-08 | 33 |
| Eastern Mediterranean | 6.054286e-05 | 3.0100e-09 | 21 |
| Western Pacific | 1.996048e-04 | 2.3783e-08 | 21 |
| South-East Asia | 1.304636e-04 | 1.4277e-08 | 11 |
| Total | 1.892087e-04 | 4.2960e-08 | 183 |

| Level | Mean | Var | N |
|---|---|---|---|
| Low-Income | 3.269452e-05 | 6.1730e-09 | 31 |
| Lower-middle-income | 7.809935e-05 | 9.5380e-09 | 46 |
| Upper-middle-income | 1.894604e-04 | 2.9307e-08 | 53 |
| High-income | 3.769377e-04 | 4.7794e-08 | 53 |
| Total | 1.892087e-04 | 4.2960e-08 | 183 |

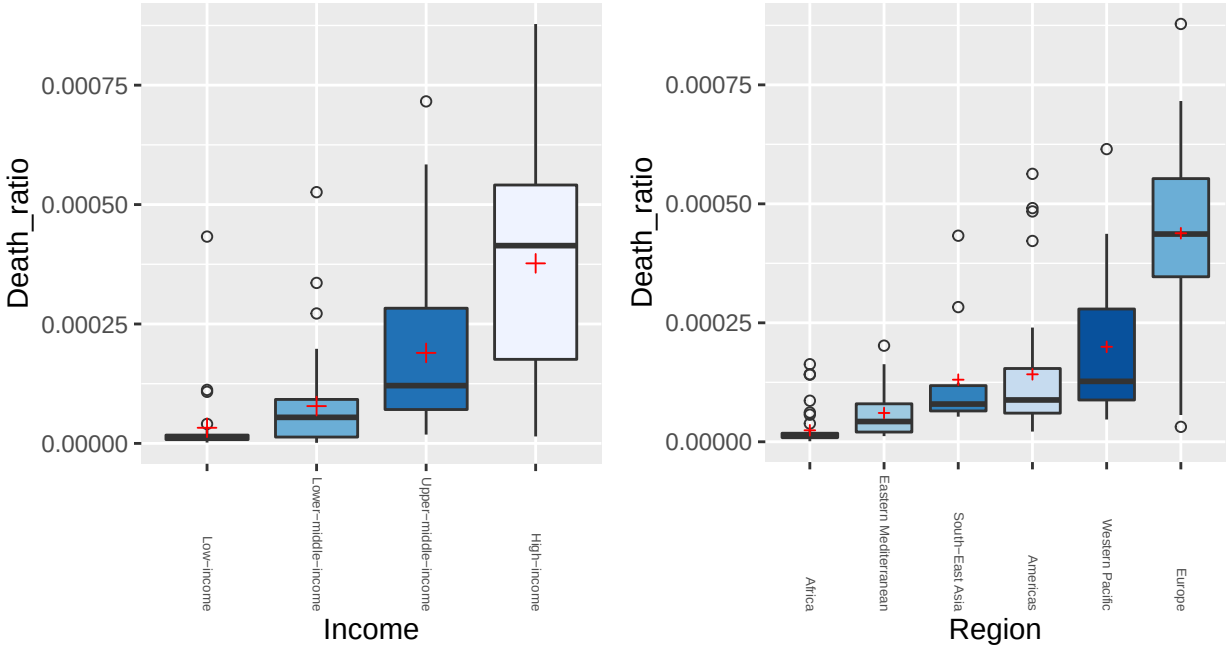


Figure 3: Boxplots of death ratio for the different levels of the categoriacl variables. On the left: Income. On the right: Regions. The red cross is the mean value for each level

- Sum of Squares Within : $SSW = \sum(y - \hat{y})^2$ which is the variation among countries within each group
- Sum of Squares Between : $SSB = \sum(\hat{y} - \bar{y})^2$ which is the variation between the groups where $\hat{y}$ is the mean of $y$'s group and $\bar{y}$ is the general mean of all $y$.

Using those measures we obtain the F-statistic as $F = \frac{SSW/df_W}{SSB/df_B}$ where $df_W = N - k$ and $df_B = k - 1$ where N is the number of observation (183 in our case) and k is the number of groups (4 in the Income case and 6 in Region's one). We used this statistic in order to test if the difference in mean between the different groups is significant or not looking at the corresponding pvalue.

Table 2: Anova Summary of Income (on the left) and of Regions (on the right)

| SSW | SSB | F statistic | P-value |
|---|---|---|---|
| 4.623625e-06 | 3.19512e-06 | 41.23218 | 2.627025e-20 |

| SSW | SSB | F statistic | P-value |
|---|---|---|---|
| 2.976671e-06 | 4.842075e-06 | 57.58428 | 2.40305e-35 |

From the result in Table 2 we can conclude that the number of deaths caused by lung cancer, for different groups of `Region` or `Income`, are significantly different in mean.

## 1.4 Issues and quality of data

The data are valid and reliable because they came from an important institutional source that is World Health Organization and they are suitable to address our aim and the size of the dataset is enough for it. The problem is that some countries didn't provide all the information needed and some values seem to be not 100% accurate, therefore there might be some

measurement problems. The problems we faced during our analysis are mainly related to the fact that the values of the variable of interest are very small, less than 0.0001, and the variable presents a bimodal distribution due to the richest countries, those with an high level of income.

## 2 Missing values

Our dataset contains 39 `NA` values in the variable `Tobacco`, for that reason we decided to estimate those values using a subset of the entire dataset in order to use a multivariate normal model for the imputation of the missing data, using Gibbs Sampler. The goal is to sample the missing values of each data point independently.

From now on the variables `Government_expenditure`, `Domestic_expenditure`, `Age_mean` and `Pollution` will be replaced by their logarithm because we want to assume that these variables are normally distributed. The graph in Figure 4 represents the variables that will form the dataset subset `Y`, whose distribution, can be assimilated to a multivariate normal.

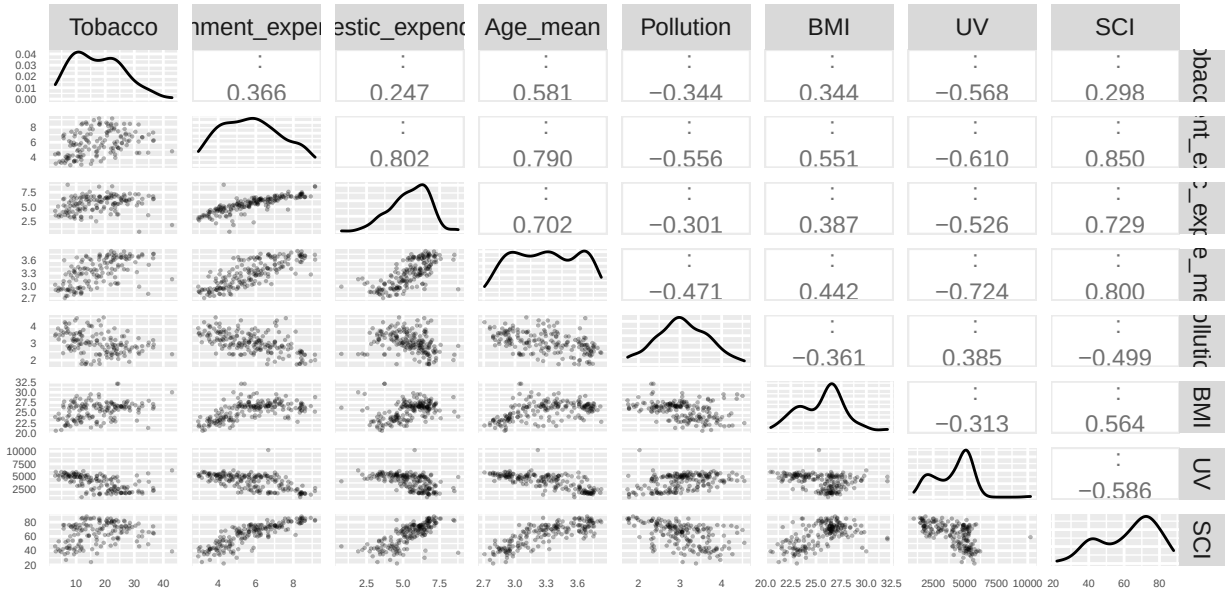


Figure 4: Representation of the quantitative variables that can be assumed to have a normal distribution. In the lower part of the graph there are the scatterplots of the variables considered in pairs. On the diagonal there are the densities of the variables. On the upper part there are the values of the correlation among the variables

Consider the dataset `Y` (made up with the variables normal distributed of our dataset) with missing values and a corresponding matrix `O` which contains the value 1 if the corresponding element in `Y` exists or 0 elsewhere. A Gibbs sampler is then implemented where the matrix `Y` is not a full matrix.

Assume that `Y` is formed by two parts $Y = (Y_{obs}, Y_{miss})$ and consider the following Bayesian model:

$$Y_1, \dots, Y_n | \theta, \Sigma \sim N_p(\theta, \Sigma)$$

$$\theta \sim N_p(\mu_0, \Lambda_0)$$

$$\Sigma \sim Inv - Wishart(\nu_0, \S_0)$$

Whose full conditionals distributions are known, then we can use Gibbs Sampler in order to estimate the posterior distribution on $Y_m iss$. First we sample values $\theta^{(s)}$ and $\Sigma^{(s)}$ and then we can sample from $Y^{(s)}_{miss} \sim p(Y_{miss}|Y_{obs}, \theta^{(s)}, \Sigma^{(s)})$ In order to do that consider as variable $a$ the index of the observed values for the i-th data, and as variable $b$ the index of the missing values for the i-th data.

Our analysis focuses on the study of the joint posterior distribution of the parameters and the unobserved quantities. So, we proceed calculating $y_{[b]} \mid y_{[a]}$ , $\theta$, $\Sigma \sim N(\theta_{b|a}, \Sigma_{b|a})$ as follows:

$$\theta_{b|a} = \theta_b + \Sigma_{[b,a]} + (\Sigma_{[a,a]})^{-1}(y_{[a]} - \theta_{[a]})$$

$$\Sigma_{b|a} = \Sigma_{[b,b]} - \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}\Sigma_{[a,b]}$$

The initial value used for $\theta$ is the mean of the data and as initial value of $\Sigma$ is used the covariance matrix of the data.

Using Gibbs sampling, after sampling the values of $\theta^{(s)}$ and of $\Sigma^{(s)}$, it is possible to sample from the distribution of the unobserved quantities conditionally on the parameters and the observed quantities. Then, we create a function to perform Bayesian analysis and we use as input for the missing data the values calculated above. The following steps are:

1) Calculate the full conditional of $\theta \mid \Sigma$, computing first $\mu_n$ and $\Lambda_n$ using the imputed full dataset

2) Calculate the full conditional of $\Sigma \mid \theta$ using the imputed dataset, computing first the value of $\nu_n$ and of $S_n$, first centering the data according to the current value of $\theta$, then computing the residual sum of square

3) Imputed the data again and update the values

```
## The Gibbs function
Gibbs_multi_norm_miss <- function(G,burnin, thin, in_theta,in_Sigma,in_Y, mu_0,
                                  Lam_0,nu_0,S_0,Y,O) {
  p <- ncol(Y); n <- nrow(Y)
  iterations <- burnin + thin * G ### Compute the iterations
  g <- 1
  theta <- matrix(nrow = G, ncol = p); Sigma <- array(dim = c(G, p, p)) ## Define the output object
  Y_imputed <- matrix(nrow = G, ncol = sum(O == 0)) ## The imputed data
  ## Quantity of interest
  Lam_0m1 <- solve(Lam_0)
  current_theta <- in_theta ## the current state of the chain
  current_Sigma <- in_Sigma
  current_Y <- in_Y ## matrix without NA (i.e. full)
  for (iter in 1:iterations) {
    #### Step 1
    ybar <- apply(current_Y, 2, mean); Sigm1 <- solve(current_Sigma)
    Lam_n <- solve(Lam_0m1 + n * Sigm1); mu_n <- Lam_n %*% (Lam_0m1 %*% mu_0 + n * Sigm1 %*% ybar)
    current_theta <- as.vector(rmvnorm(1, mean = mu_n, sig = Lam_n))
    ### Step 2
    nu_n <- nu_0 + n; Z <- scale(current_Y, center = current_theta, scale = rep(1, p))
    S_theta <- t(Z) %*% (Z); S_n <- S_0 + S_theta
    Omega <- rWishart(1, df = nu_n, Sigma = solve(S_n))
    current_Sigma <- solve(Omega[, , 1])
    ### Step 3 repeat the process above
    for (i in 1:n) {
      if (all(O[i,] == 1)) {
        next}
      oi = O[i,]; a = (oi == 1); b = (oi == 0)
      iSa = solve(current_Sigma[a, a]); beta.j = current_Sigma[b, a] %*% iSa
      Sigma.j = current_Sigma[b, b] - beta.j %*% current_Sigma[a, b]; yi = current_Y[i,]
      theta.j = current_theta[b] + beta.j %*% t(yi[a] - current_theta[a])
      current_Y[i, b] = as.vector(rmvnorm(1, theta.j, Sigma.j))
      while (current_Y[i, b] <= 0) {
        current_Y[i, b] = as.vector(rmvnorm(1, theta.j, Sigma.j))}}
    ## In the output objects I save only the state reached after a burn-in
    ## Moreover to decrease the correlation between subsequent state we can thin out the chain
    if ((iter > burnin) & (iter %% thin == 0)) {
      theta[g, ] = current_theta
      Sigma[g, , ] = current_Sigma
      Y_imputed = current_Y[O == 0]
      g = g + 1}}
  return(list(theta = theta, Sigma = Sigma, Y_imputed = Y_imputed))}
mu_0 <- apply(Y, 2, mean, na.rm = T);
sd_0 <- (mu_0 / 2); Lam_0 <- matrix (.1, p, p); diag(Lam_0) <- 1
Lam_0 <- Lam_0 * outer(sd_0 , sd_0)
nu_0 <- p + 2; S_0 <- Lam_0
#out=Gibbs_multi_norm_miss(5000,1000,5,in_theta = in_theta,in_Sigma = in_Sigma,
```

```
#                     in_Y=in_Y,mu_0=mu_0,Lam_0=Lam_0,nu_0=nu_0,
#                     S_0=S_0,Y=Y,O=O)
```

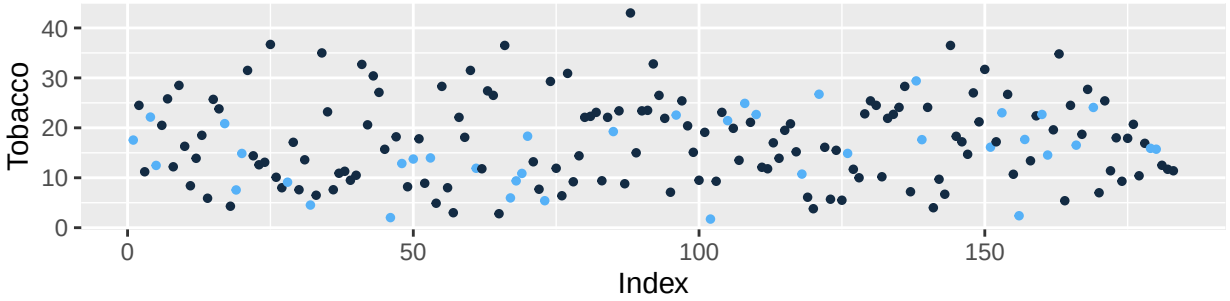


Figure 5: Scatterplot of the variable Tobacco. Dark blue dots are observed values. Sky blue dots are imputated ones

We checked the convergence and dependance of the MCMC through trace plots and acf plots, and we concluded that everything seems to be good, there is good mixing and no autocorrelation among the chain.

# 3 The model

The model that better describes the response `Death Ratio` is a Bayesian Beta Regression Model defined as follow:

$$\begin{aligned} Y_i \mid \mu_i, \phi &\sim Beta(\mu_i\phi, \phi(1-\mu_i)) \\ logit(\mu_i) &= x_i^T\beta \\ \phi &\sim Gamma(\alpha, \beta) \end{aligned} \tag{1}$$

with $\beta = (\beta_1, ..., \beta_p)$ and $\beta_j \sim N(0, \sigma^2_{\beta_j})$ $\quad j = 1, ..., p \quad i = 1, ..., n.$

The choice of a beta distribution is optimal when the data is continuous and restricted to the interval (0,1) as in this case where the response is a ratio. As it is possible to observe from Figure 1, the response variable has really small positive values and its density distribution is bimodal and not symmetric. The Beta distribution seems to be optimal for this case because it is characterizes by a high level of flexibility in terms of the assortment of density shapes that can be accommodated (Adam J. Branscum 2007).

Equation (1) is made up of the beta prior that has a specific parametrization to incorporate in an easiest way the covariate information. The parameters are: $a = \mu_i\phi$ and $b = \phi(1-\mu_i))$. $\mu$ is defined as $\mu = \frac{a}{a+b}$ and $\phi = a + b$ is a parameter related to the variance of the distribution that is equal to $\sigma^2 = \frac{\mu(1-\mu)}{\phi+1}$. As a consequence, variance increases as $\phi$ decreases. Another important result is the fact that the variance previously defined is not constant because it depends on $\mu_i$ and in this way the model is even more appropriate to describes not-symmetric data. The link function is the logit link, therefore $\mu_i = \frac{exp(x_i^T\beta)}{1+exp(x_i^T\beta)}$. Finally, $\phi$ is distributed as a Gamma density with parameters $\alpha$ and $\beta$ because it is always positive.

## 3.1 Parameters Choice

The parameter $\phi$ is distributed as a $Gamma(10, 0.1)$ where shape = 10 and rate = 0.1 with $E(X) = \frac{shape}{rate}$ and $Var(X) = \frac{shape}{rate^2}$. The value are chosen so that the variance is relatively big and therefore the distribution of $\phi$ is weakly informative. At the same time, the mean has to be evaluated in relation to the variance of the beta distribution previously defined. It is required to be small so that the variance of the beta will be relatively large compared to the average values of the response, but at the same time it should be quite large because the order of magnitude of the response variable is really small ($10^{-7}$). Finally, we assume a simple non-informative prior for the coefficients $\beta_j$: centered in 0 and with a large variance (precision equal to 0.1).

# 4 Model Selection

The best models among the $2^p$ models can be selected with spike and slab model defined as follow:

$$\begin{aligned} \beta_j \mid \sigma_j^2 &\sim N(0, \sigma_j^2) \\ \sigma_j^2 \mid c_j, \tau_j, \gamma_j &\sim (1-\gamma_j)\delta_{\tau_j^2} + \gamma_j \, \delta_{\tau_j^2 \; c_j^2} \\ \gamma_j \mid \theta &\sim Bern(\theta_j) \\ \theta_j &\sim Unif(0,1) \end{aligned} \tag{2}$$

The peculiarity of this model is the presence of a variable selection prior for $\beta_j$ that is a combination between a slab component and an almost spike component. This spike and slab priors "induce a positive prior probability on the hypothesis $H_0 : \beta_j = 0$" (Rockova and George 2018). The quasi spike component is defined as a Gaussian with variance $\tau^2 > 0$ that is a really small value. It is defined as: $\tau = \frac{k}{\sqrt{2\frac{log(c)c^2}{c^2-1}}}$. The slab component is a normal with large variance defined as: $c^2\tau^2 > 0$. The value $\pm k$ defines the points where the two densities intersect and also the distance $[-k, +k]$ needs to be small because it represents an approximation of 0. Considering all these elements the values chosen for c and k are: $c = 200$ and $k = 0.001$ and, as a consequence, $\tau^2 = 2.35918110^{-6}$ and the variance of the slab component is equal to 0.003774689.

To be sure to reach the convergence of the chain the value of the iterations (G) has been set equal to 120000, while the burn-in is set equal to 20000. At the same time, to reduce auto correlation within the chain the value of the thinning is set equal to 5. In this way 100000 iterations are saved. The chains created by JAGS are 5 and each of them has a different initial value for $\phi$: 1, 1000, 2000, 3000 and 4000. The initial values for $\beta_j$ and for $\gamma$ are kept constant for all the chains equal to 0.

```
ssvs_model <- function() {
  for (i in 1:n) {
    logit(mu[i]) <- beta0 + inprod(X[i,], beta[])
    Y[i] ~ dbeta(phi * mu[i], phi * (1 - mu[i]))
  }
  ## Parameters that do not depend on the covariates
  beta0 ~ dnorm(0, 0.1) #Intercept
  phi ~ dgamma(10, 0.1) # Parameter phi
  for (j in 1:p) {
    sig2[j] <-
      equals(gamma[j], 0) * var_spike + equals(gamma[j], 1) * var_slab
    prec[j] <- 1 / sig2[j]; beta[j] ~ dnorm(0, prec[j])
    gamma[j] ~ dbern(theta[j])}; var_spike <- tau2; var_slab  <- cc * tau2
  for (j in 1:p) {
    theta[j] ~ dunif(0, 1)}}; tau2 <- tau_ss ^ 2; cc <- c_ss ^ 2
data_jags <-list("n" = n, "p" = p, "Y" = Y, "X" = as.matrix(X), "tau2" = tau2, "cc" = cc)
```

```
####Posterior parameters JAGS  has to save
params <- c("beta0", "beta", "gamma", "phi")
## Some other information for the MCMC algorithm
burn <- 20000; thin <- 5; nit <- burn + 100000
set.seed(123)
# spsl <- jags(data = data_jags, inits = inits, parameters.to.save = params,
# model.file = ssvs_model, n.chains = 5, n.iter = nit, n.thin = thin, n.burnin = burn, DIC = T)
```

## 4.1 Checking convergence

From Figure 6 it is possible to perform a qualitative analysis of the convergence of the posterior chain with the trace-plot of the deviance and the ergodic plot that expresses stationarity. Also from the auto-correlation plot it seems that the auto-correlation is really low. Further checks can be done by looking at the Gelman diagnostic output below that diagnoses convergence because all the values of the upper limit are close to 1.

Table 3: Gelman Diagnostic

| | Point.Est. | Upper CI | Var | Point.Est. | Upper CI | Var | Point.Est. | Upper CI |
|---|---|---|---|---|---|---|---|---|
| beta[1] | 1.0076 | 1.0173 | beta[15] | 1.0004 | 1.0005 | gamma[8] | 1.0177 | 1.0437 |
| beta[2] | 1.0029 | 1.0069 | beta[16] | 1.0002 | 1.0004 | gamma[9] | 1.0009 | 1.0021 |
| beta[3] | 1.0091 | 1.014 | beta[17] | 1.0001 | 1.0002 | gamma[10] | 1.0034 | 1.009 |
| beta[4] | 1.0007 | 1.001 | beta[18] | 1 | 1.0001 | gamma[11] | 1.0034 | 1.0076 |
| beta[5] | 1.0017 | 1.0038 | beta[19] | 1.0001 | 1.0002 | gamma[12] | 1.0004 | 1.0012 |
| beta[6] | 1.0105 | 1.0234 | beta0 | 1.0113 | 1.0293 | gamma[13] | 1.0008 | 1.0023 |
| beta[7] | 1.0043 | 1.0115 | deviance | 1.0005 | 1.0013 | gamma[14] | 1.0007 | 1.0021 |
| beta[8] | 1.0186 | 1.0394 | gamma[1] | 1.0057 | 1.0148 | gamma[15] | 1.0014 | 1.0038 |
| beta[9] | 1.0018 | 1.0034 | gamma[2] | 1.0014 | 1.004 | gamma[16] | 1.0004 | 1.0011 |
| beta[10] | 1.004 | 1.0075 | gamma[3] | 1.0057 | 1.0158 | gamma[17] | 1.0006 | 1.0017 |
| beta[11] | 1.0027 | 1.0047 | gamma[4] | 1.0007 | 1.002 | gamma[18] | 1.0004 | 1.0012 |
| beta[12] | 1.0004 | 1.0009 | gamma[5] | 1.0027 | 1.0076 | gamma[19] | 1.0009 | 1.0027 |
| beta[13] | 1.0007 | 1.0008 | gamma[6] | 1.0068 | 1.0184 | r | 1.0002 | 1.0005 |
| beta[14] | 1.0002 | 1.0004 | gamma[7] | 1.0012 | 1.0013 | | | |

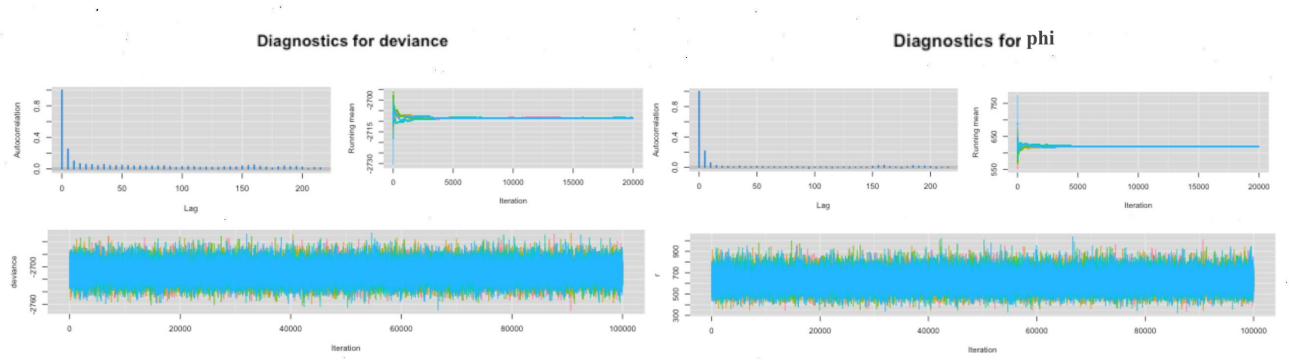


Figure 6: On the left: Autocorrelation plot, ergodic plot and traceplot of the deviance. On the right: Autocorrelation plot, ergodic plot and traceplot of the parameter phi

## 4.2 Highest posterior density

This method considers the posterior probability of each model and select as the best model the one with the highest posterior probability.First of all we selected all the models visited by the MCMC. Then, we take into consideration all the unique models among the visited ones and compute their posterior frequencies. Finally, the model with the highest frequency is selected and is the one with highest posterior density. The model that result is the one with the variables: `Tobacco`, `Government_expenditure`, `Pollution`, `IncomeLower-middle-income`, `IncomeUpper-middle-income` and `RegionEurope`.

## 4.3 Median probability model

This method will select as significant variables the ones such that the posterior probability of inclusion of the coefficient $\beta_j$ is higher than 0.5. To reach this aim, the posterior chain of gamma has been extracted. At this point we compute the sample mean of the posterior chain for each value of gamma is computed obtaining 19 values of the posterior mean corresponding to each variables. The variables are then selected if their posterior mean is higher than 0.5. The variables selected with this method are: `Government_expenditure`, `Pollution`, `RegionAmericas`, `RegionEurope` and `RegionSouth-East Asia`.

## 4.4 Hard selection Shrinkage

Another method that can be used is the hard selection shrinkage that includes in the best model only the variables for which the marginal posterior credible interval does not contain the value 0. In particular, we started by extracting the posterior chain of the beta variables and then we proceed computing the sample mean and the standard deviation, column by column for each beta. Thus, we can calculate the marginal posterior credible interval and we consider as significant the variables whose range does not include 0.

In Figure 7 are shown the posterior intervals that contain or not the 0. All the intervals contain 0, with the exception of X7 which is `UV`, even if there is a bit of uncertainty.

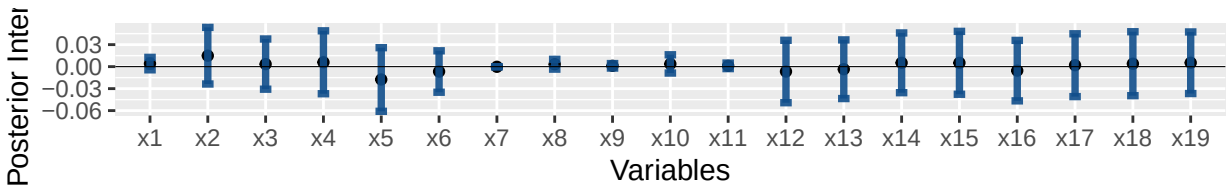


Figure 7: Decision intervals for Hard Shrinkage

# 5 Final Model

At this point we focus out analysis on the model with variables selected with the HPD method. Considering the model at Equation 1, the linear predictor will be: $logit(\mu_i) = \beta_0 + \beta_1\ x_{Tobacco} + \beta_2\ x_{Government\_expenditure} + \beta_3\ x_{Pollution} + \beta_4\ x_{IncomeLow} + \beta_5\ x_{IncomeLower-middle} + \beta_6\ x_{IncomeUpper-middle} + \beta_7\ x_{RegionAmericas} + \beta_8 x_{RegionEasternMediterranean} + \beta_9\ x_{RegionEurope} + \beta_{10}\ x_{RegionSouth-EastAsia} + \beta_{11}\ x_{RegionWesternPacific}$

The model is implemented using 120000 iterations, with 20000 of burn-in and thinning equal to 5. The number of chains are still 5 with the same initial values used in the spike and slab model.

```
model_string <- function() {for (i in 1:n) {
    Y[i] ~ dbeta(phi * mu[i], phi * (1 - mu[i]))
    logit(mu[i]) <- inprod(X[i, ], beta[])}
  for (j in 1:p) {
    beta[j] ~ dnorm(0, 0.1) }
  phi ~ dgamma(10, 0.1)}
burn <- 20000; nit <- burn + 100000; thin <- 5
# out1 <- jags(model.file=model_string, data = data1,
#               inits = inits, parameters.to.save = params, n.iter = nit,
#               n.burnin = burn, n.thin = thin,n.chains= 5)
```

Table 4: Summary

| | mean | sd | X2.5. | X25. | X50. | X75. | X97.5. | n.eff |
|---|---|---|---|---|---|---|---|---|
| beta[1] | -7.5984 | 1.0737 | -9.7142 | -8.3203 | -7.6039 | -6.8695 | -5.4980 | 9000 |
| beta[2] | 0.0107 | 0.0105 | -0.0102 | 0.0036 | 0.0108 | 0.0179 | 0.0310 | 19000 |
| beta[3] | 0.0377 | 0.1120 | -0.1826 | -0.0377 | 0.0378 | 0.1146 | 0.2548 | 7800 |
| beta[4] | -0.1153 | 0.1654 | -0.4413 | -0.2264 | -0.1154 | -0.0032 | 0.2079 | 23000 |
| beta[5] | -0.2636 | 0.4528 | -1.1524 | -0.5689 | -0.2628 | 0.0425 | 0.6236 | 9400 |
| beta[6] | -0.2331 | 0.3688 | -0.9551 | -0.4830 | -0.2348 | 0.0152 | 0.4922 | 9000 |
| beta[7] | -0.0816 | 0.2572 | -0.5818 | -0.2567 | -0.0839 | 0.0914 | 0.4260 | 9100 |
| beta[8] | 0.1888 | 0.2642 | -0.3270 | 0.0109 | 0.1887 | 0.3668 | 0.7066 | 79000 |
| beta[9] | 0.0936 | 0.2947 | -0.4979 | -0.1018 | 0.0983 | 0.2928 | 0.6592 | 51000 |
| beta[10] | 0.5289 | 0.2875 | -0.0335 | 0.3364 | 0.5285 | 0.7219 | 1.0940 | 56000 |
| beta[11] | 0.3169 | 0.3356 | -0.3768 | 0.0996 | 0.3280 | 0.5475 | 0.9451 | 100000 |
| beta[12] | 0.2616 | 0.3070 | -0.3471 | 0.0574 | 0.2630 | 0.4690 | 0.8570 | 65000 |
| deviance | -2703.4631 | 15.2909 | -2732.7434 | -2713.8114 | -2703.7426 | -2693.3068 | -2672.6958 | 53000 |
| phi | 623.0315 | 83.8488 | 468.3747 | 564.8502 | 619.7136 | 677.4474 | 796.9911 | 100000 |

In the upper part of Figure 8 are illustrated the posterior density of the intercept (on the left) and of the coefficient $\beta_2$ corresponding to the variable `Tobacco`. The mean of $\beta_1$ is a negative value equal to -7.5984 and its entire credible interval includes negative values. In contrast, the mean coefficient of the variable `Tobacco` is a positive value equal to 0.0107. In this case the credible interval always includes the 0, therefore some possible values of the coefficient are negative. The interpretation of this coefficient is that an increase of 1% in the cigarette smoking prevalence, ceteris paribus, has a positive increase on the logit of the mean of the death ration of a country. As it is possible to notice in the lower part of Figure 8, the convergence value of $\phi$ is very high, it is equal to 623.0315. As represented in the graph the 95% values generated by the posterior density of $\phi$ lie within the interval 468.3747 and the third is equal to 796.9911.

Concerning the coefficients of the qualitative variables, $\beta_{5:7}$ are the coefficients of the levels of Income, where the baseline is the level High-income. On the other hand, $\beta_{8:12}$ are the coefficients of the levels of Region, where the baseline is the region Africa. In Figure 9 we can notice that the level of Low-income has the smallest impact in the logit of the mean of the

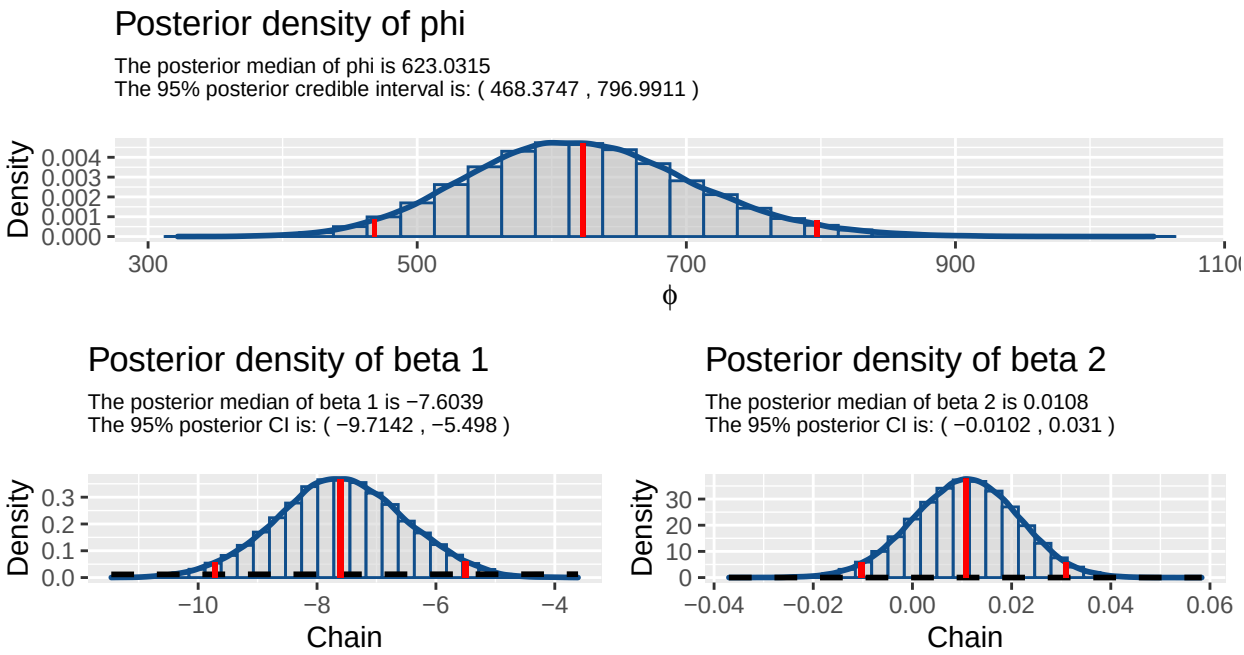


Figure 8: On the upper part: Posterior density of phi. On the lower part: Posterior density for coefficients 1 and 2. The continous red line represents the mean value of the posterior density. The dotted red lines indicate the first and third quantile of the distribution. The blue line is the posterior density. The black dotted line is the prior distribution.

mortality ratio. The others levels of `Income` have higher but negative coefficients, anyway they do not differ substantially. In conclusion, we can say that higher is the Income per country, higher is the probability that that country has an higher value of the logit of the mean of the ratio of deaths caused by lungs cancer. Moreover, in Figure 9 it's evident that the region Europe has the highest impact on the logit of the mean of the mortality ratio, while the baseline, which is Africa, has the smallest. There is an upward trend that starts with the baseline in 0 and increases. However, all coefficients, except a small part of the one corresponding to `Eastern Mediterranean`, have values whose 95% credible interval is positive but close to zero. In conclusion, the regions ordered depending on their probability of having, for the logit of the mean, an higher ratio of deaths caused by lungs cancer, are: `Africa`, `Eastern Mediterranean`, `Americas`, `Western Pacific`, `Soth-East Asia` and `Europe` .

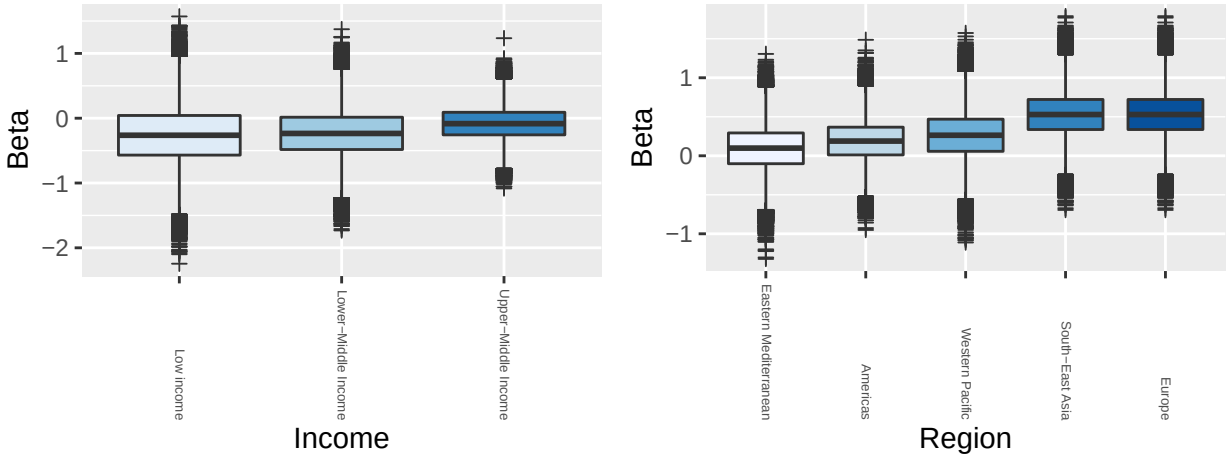


Figure 9: Boxplots of the values of betas for the categorical variables

## 5.1 Checking convergence

An important consideration, in order to verify the success of the MCMC, is to check the convergence and dependence through all the trace-plots, ergodic plots and acf plots. In Figure 10 are reported the graphs for the `deviance`, for the parameter `phi` and for the coefficient $\beta_{12}$. The plots seem to have a good mixing: the chains seem to be not auto-correlated and they

Table 5: Gelman Diagnostic

| | Point.Est. | Upper CI | Var | Point.Est. | Upper CI |
|---|---|---|---|---|---|
| beta[1] | 1.0068 | 1.0173 | beta[8] | 1.0001 | 1.0003 |
| beta[2] | 1.0004 | 1.0012 | beta[9] | 1.0001 | 1.0004 |
| beta[3] | 1.0059 | 1.0154 | beta[10] | 1.0004 | 1.0009 |
| beta[4] | 1.0024 | 1.0055 | beta[11] | 1.0001 | 1.0002 |
| beta[5] | 1.0054 | 1.0138 | beta[12] | 1.0006 | 1.0016 |
| beta[6] | 1.0049 | 1.0122 | deviance | 1.0005 | 1.0015 |
| beta[7] | 1.0036 | 1.0094 | phi | 1.0001 | 1.0004 |

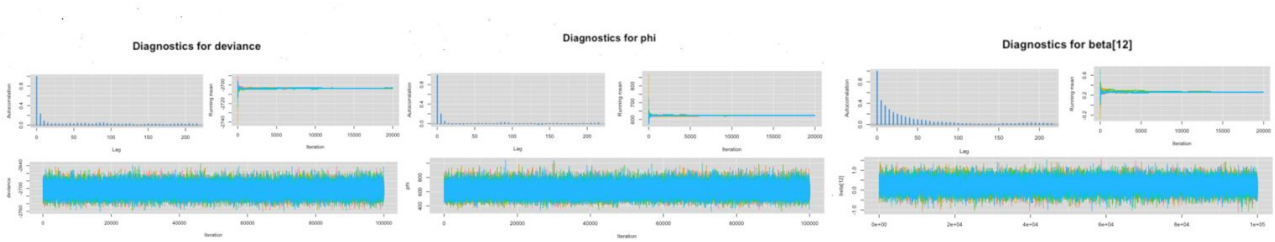


Figure 10: On the left: Autocorrelation plot, ergodic plot and traceplot of the deviance. On the center: Autocorrelation plot, ergodic plot and traceplot of the parameter phi. On the right: Autocorrelation plot, ergodic plot and traceplot of beta 12

converge apparently at the same value. For others coefficients beta the convergence was still good, while the autocorrelation wasn't really small, in particular with low values of lag. In order to have a quantitative analysis for the convergence we performed the Gelman Diagnostic (Table 5), that confirms the good results previously mentioned, since all the values of the potential scale reduction factor are very close to 1. However, as regards the auto-correlation we have computed the Effective Sample Size, that we can see in the last column of the summary. We notice that there is no substantial discrepancy between the number of iterations and the ESS. Therefore, also the auto-correlation shouldn't be a problem of our chains.

## 5.2 Model Checking:

### 5.2.1 Posterior predictive distributions

A Bayesian approach to verify the adequacy of the model is through the posterior predictive distributions, the idea is to use the observed values a statistic (we tried with the mean, the first and the third quantile) and check its plausibility against the corresponding posterior predictive distribution.

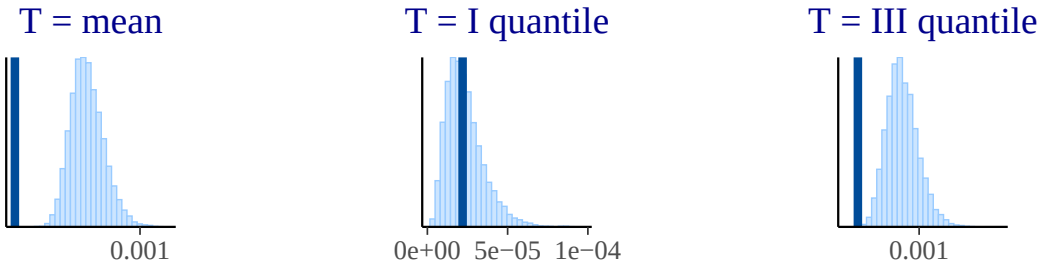


Figure 11: Model checking. First plot: posterior distribution of the mean with vertical line representing the observed value of the mean. Second plot: posterior distribution of the probability of the death ratio of being less than the first quantile of the observed values. The vertical line representing the observed value of this statistic. Third plot: considering the probability of the death ratio of being larger than the third quantile of the observed values. The vertical line represents the observed value of this statistic

Observing Figure 11 the conclusion could be that the model is not good if we are interested in the average of the ratio of cancer deaths, we can see that the observed value (blue line) lies in the tails of the distribution, that means that our model performs poorly in estimating this statistic. However, if we are interested in the first quantile (25%) we can see from Figure 11 that the observed value lies inside the distribution, so the value is plausible and the model is useful for that statistic.

Finally, the model is not useful and not good with regard to the third quantile because the observed value, as for the mean, lies completely out of the distribution.

### 5.2.2 Bayesian residuals and model adequacy

Another approach in order to verify if the model is suitable for our data is using the Bayesian residuals, which is the standardization of the observed response minus the predicted one ($R_i = \frac{y_i - \hat{Y}_i^*}{\sigma}$). A model can be considered good if the $\hat{Y}^*$'s are quite similar to the $y$ values, namely if the Bayesian residuals are around 0. Though, we can see in Figure 12 that the predicted Y doesn't coincide with the real values, in particular they are overestimated, while in Figure 12 we notice that Bayesian residuals are quite around 0.

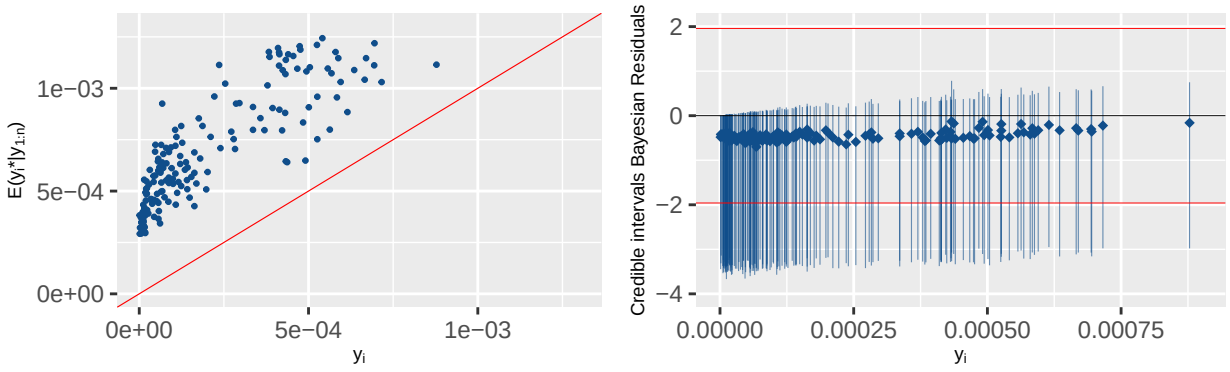


Figure 12: Model adequacy

# 6 Best models comparison

Despite the method established as the choice of variables was the HPD one we decided to implement a Beta regression model also using the variables of the other two methods illustrated in paragraphs 4.3 and 4.4. The second model was constructed using as covariates the variables `Government_expenditure`, `Pollution` and `Region`, as established for the MDM criteria. The third model has as only covariate the variable `UV`, as established using the HS criteria. We repeated the same procedure we used for the HPD model. We then checked the convergence and the auto-correlation of the chains of both the models.

The three best models obtained with these approaches can be compared with the Deviance Information Criterion. This method takes into account both the goodness of fit and the complexity of the model: $DIC = \bar{D} + P_D$. $\bar{D}$ is the posterior mean deviance that is equal to: $\bar{D} = \frac{1}{G}\sum_{g=1}^{G} D(\theta_g)$ and measures the goodness of fit. $P_D$ is a measure of model complexity and is based on the asymptotic representation of the likelihood and it is: $P_D = \bar{D} - D(\hat{\theta})$. $D(\hat{\theta})$ is the deviance of the posterior mean of $\theta$.

As it is possible to observe from Table 6, the DIC of the models are not really different, in fact the difference between the first and the third model is 19.268 and between the first and the second is even less. Moreover, most of this difference is caused by the higher complexity of the first model that considers 11 regressors, while the third model is built considering only 1 regressor.

Table 6: DIC

| | HPD Model | MPM Model | HS Model |
|---|---|---|---|
| DIC | -2586.562 | -2594.681 | -2605.830 |
| pD | 116.901 | 110.967 | 98.135 |
| Deviance | -2703.463 | -2705.648 | -2703.965 |

# References


Adam J. Branscum, Mark C. Thurmond, Wesley O. Johnson. 2007. "Bayesian Beta Regression: Applications to Household Expenditure Data and Genetic Distance Between Foot-and-Mouth Disease Viruses." *Australian & New Zealand Journal of Statistics* Volume 49 (Issue 3): p. 287–301.

Cepeda-Cuervo, Edilberto, Daniel Jaimes, Margarita Marı'n, and Javier Rojas. 2015. "Bayesian Beta Regression with Bayesianbetareg R-Package." *Computational Statistics* 31 (1). Springer Science; Business Media LLC: 165–87. https://doi.org/10.1007/s00180-015-0591-9.

Hoff, Peter D. 2009. *A First Course in Bayesian Statistical Methods*. Springer New York. https://doi.org/10.1007/978-0-387-92407-6.

Kruschke, John K. 2015. *Doing Bayesian Data Analysis*. 2nd ed. Elsevier.

Rockova, Veronika, and Edward I. George. 2018. "The Spike-and-Slab LASSO." *Journal of the American Statistical Association* 113 (521). Informa UK Limited: 431–44. https://doi.org/10.1080/01621459.2016.1260469.