Table of contents I

Sustainable Development Goal


OBJETIVOS DE DESARROLLO SOSTENIBLE
1 FIN DE LA POBREZA
2 HAMBRE CERO
3 SALUD Y BIENESTAR
4 EDUCACIÓN DE CALIDAD
5 IGUALDAD DE GÉNERO
6 AGUA LIMPIA Y SANEAMIENTO
7 ENERGÍA ASEQUIBLE Y NO CONTAMINANTE
8 TRABAJO DECENTE Y CRECIMIENTO ECONÓMICO
9 INDUSTRIA, INNOVACIÓN E INFRAESTRUCTURA
10 REDUCCIÓN DE LAS DESIGUALDADES
11 CIUDADES Y COMUNIDADES SOSTENIBLES
12 PRODUCCIÓN Y CONSUMO RESPONSABLES
13 ACCIÓN POR EL CLIMA
14 VIDA SUBMARINA
15 VIDA DE ECOSISTEMAS TERRESTRES
16 PAZ, JUSTICIA E INSTITUCIONES SÓLIDAS
17 ALIANZAS PARA LOGRAR LOS OBJETIVOS
OBJETIVOS DE DESARROLLO SOSTENIBLE

Some targets of SDG 2 (Zero Hunger)

By 2030, end hunger and ensure access to all people, particularly the poor and those in vulnerable situations, including infants under 1 year of age, to healthy, nutritious, and sufficient food all year round.

- Prevalence of undernourishment.
- Prevalence of moderate or severe food insecurity in the population, according to the Food Insecurity Experience Scale.

Some targets of SDG 8 (Decent Work)

By 2030, achieve full and productive employment and decent work for all women and men, including youth and persons with disabilities, and equal pay for equal work.

- Unemployment rate, disaggregated by gender, age, and persons with disabilities.

Fundamental Principle of Data Disaggregation

Indicators of the Sustainable Development Goals should be disaggregated, whenever relevant, by income, gender, age, race, ethnicity, migratory status, disability, and geographical location, or other characteristics, in accordance with the Fundamental Principles of Official Statistics.

General Assembly Resolution - 68/261

Survey Limitations.

What is the Coefficient of Variation?

The coefficient of variation is a measure of relative error for an estimator and is defined as:

$$cve\left(\hat{\theta}\right) = \frac{SE\left(\hat{\theta}\right)}{\hat{\theta}}$$

It is often expressed as a percentage, even though it is not bounded to the right, which makes it convenient when discussing the precision of statistics derived from surveys.

Alert Standards in Some Countries (Household Surveys)

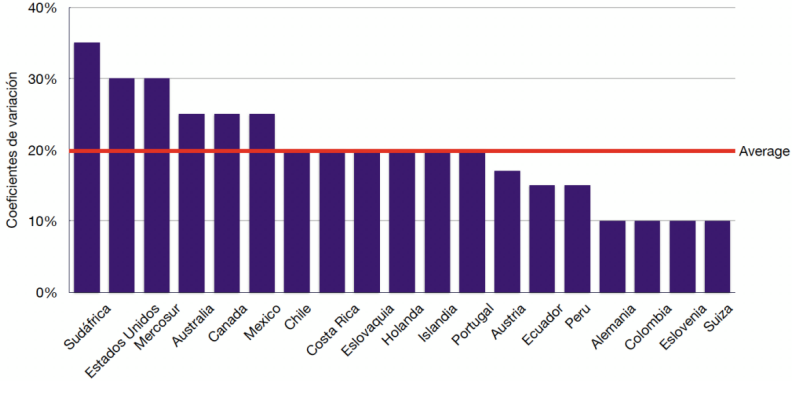


Figure 1: Alerts on Coefficients of Variation

Some Alerts Defined in the Publication

When the threshold of the coefficient of variation is exceeded, some of the following alerts may appear:

- Not published.
- Use with caution.
- Estimates require revisions, are not precise, and should be used with caution.
- Unreliable, less precise.
- Does not meet publication standards.
- With reservation, reference, questionable.
- Very random values, poor estimation.

Study Domains and Subpopulations of Interest

A survey is designed to generate accurate and reliable information within predefined study domains. However, there are subpopulations that the survey did not address in its design but for which greater precision is desired.

- Poverty incidence disaggregated by department or province (known and planned sample size).
- Unemployment rate disaggregated by gender (random but planned sample size).
- Net primary school attendance rate disaggregated by income quintiles (random sample size).

Precision of Estimators

Because a survey is a partial investigation of a finite population, it's important to know that:

- Indicators are not calculated from a survey; they are estimated using survey data.
- It is necessary to calculate the degree of error resulting from the inability to conduct a comprehensive investigation. This error is known as sampling error.
- The precision of an estimator is dependent on the confidence interval.

A narrower interval results in greater precision and, therefore, lower sampling error.

Effective Sample Size

- In household surveys with complex sampling designs, there is no sequence of variables that are independent and identically distributed.
- The sample $y_1, \ldots, y_n$ is not a vector in an n-dimensional space, where each component of the vector can vary independently.
- The final dimension of the vector $(y_1, \ldots, y_n)$ is much smaller than n, due to hierarchical sampling and the relationship between the variable of interest and primary sampling units (PSUs).

Effective Sample Size

The effective sample size is defined as follows:

$$n_{effective} = \frac{n}{Deff}$$

Where Deff is the design effect, which depends on: *1.* The average number of surveys conducted in each PSU. *2.* The correlation between the variable of interest and the same PSUs.

It can be considered that if the effective sample size is not greater than a threshold, then the figure should not be considered for publication.

Degrees of Freedom

In subpopulations, degrees of freedom are not considered fixed but rather variable.

$$df = \sum_{h=1}^{H} v_h \times (n_{Ih} - 1)$$

Note that ν_h is an indicator variable that takes the value one if stratum h contains one or more cases of the subpopulation of interest, and n_{Ih} is the number of primary sampling units (PSUs) in the stratum. In the most general case, degrees of freedom are reduced to the following expression:

```
df = #PSUs - #Stratum
```

Introduction to Bayesian Thinking.

Models of Areas with the **Tom Approach**

And you wake up one day...

- You feel a little strange and weak. You go to the doctor, and they run some tests. One of them comes back positive for a very rare disease that only affects 0.1% of the population.

Not good news.

- You go to the doctor's office and ask how specific the test is. They tell you it's very accurate; it correctly identifies 99% of the people who have the disease.

And you meet Thomas...

Esta es la información que tienes:

- $\mathrm{P(E)} = 0.001$
- $\mathrm{P}(+|\mathrm{E}) = 0.99$
- $\mathrm{P}(-\mathrm{E}) = 0.999$
- $\mathrm{Pr}(+|-\mathrm{E}) = 0.01$

Además, por el teorema de probabilidad total

$$\begin{aligned}\mathrm{P}(+) &= \mathrm{Pr(E)Pr}(+|\mathrm{E}) + \mathrm{Pr}(-\mathrm{E})\mathrm{Pr}(+|-\mathrm{E}) \\ &= 0.001 * 0.99 + 0.999 * 0.01 \\ &= 0.01098\end{aligned}$$

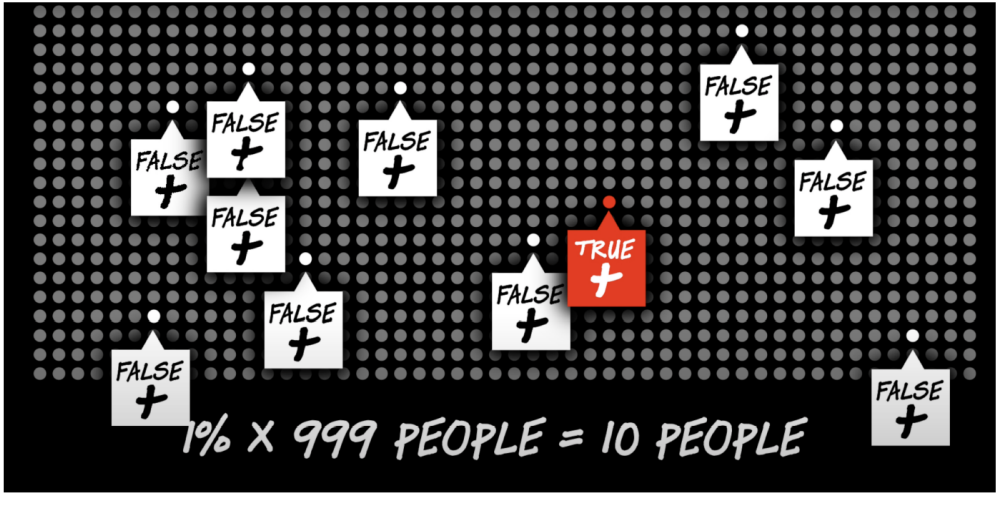
La regla de Bayes afirma lo siguiente:

$$\mathrm{Pr}(\mathrm{E}|+) = \frac{\mathrm{Pr}(+|\mathrm{E}) \times \mathrm{Pr(E)}}{\mathrm{Pr}(+)}$$

Por lo tanto:

$$\mathrm{Pr}(\mathrm{E}|+) = 0.09 \approx 9\%$$

How does it work?



How does it work?

And you seek a second opinion

- This time the doctor orders you to retake the same test... and you test positive for that disease again.
- **And you wonder again:** *What is the probability that I have this disease?*

This time, you've updated your information about $Pr(E)$ because you've tested positive on a test

$$Pr(E) = 0.09 \; And \; Pr(-E) = 0.91$$

Therefore:

$$Pr(E \mid ++) = 0.997 \approx 91\%$$

Elements of Bayes' Rule

In terms of inference for θ, it's necessary to find the distribution of the parameters conditioned on the observation of the data. To achieve this, you need to define the joint distribution of the variable of interest with the parameter vector.

$$p(\theta, Y) = p(\theta)p(Y \mid \theta)$$

- The distribution $p(\theta)$ is known as the prior distribution.
- The term $p(Y \mid \theta)$ is the sampling distribution, likelihood, or data distribution.
- The distribution of the parameter vector conditioned on the observed data is given by

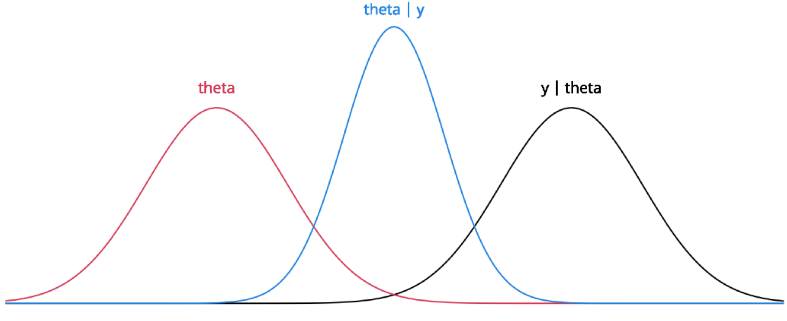
$$p(\theta \mid Y) = \frac{p(\theta, Y)}{p(Y)} = \frac{p(\theta)p(Y \mid \theta)}{p(Y)}$$

Bayes' Rule

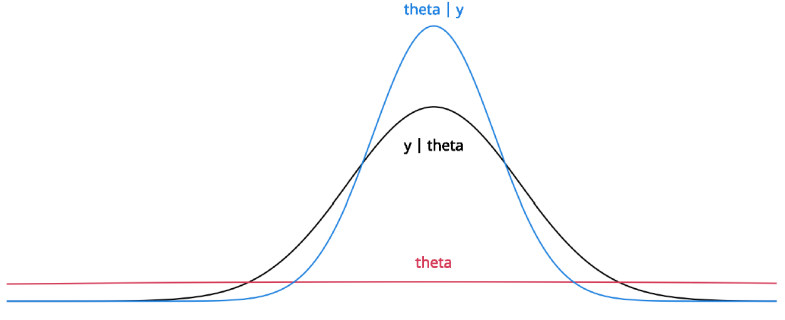
- The term $p(\theta \mid Y)$ is known as the ***posterior*** distribution.
- The denominator does not depend on the parameter vector, and considering the observed data as fixed, it corresponds to a constant and can be omitted. Therefore,

$$p(\theta \mid Y) \propto p(Y \mid \theta)p(\theta)$$

Informative Prior Distribution for θ



Non-Informative Prior Distribution for θ



Poisson Area Model

Suppose $Y = \{Y_1, \ldots, Y_n\}$ is a random sample of variables with a Poisson distribution with parameter θ. The joint distribution function or likelihood function is given by

$$p(Y \mid \theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{y_i}}{y_i!} I_{\{0,1,\ldots\}}(y_i)$$

$$= \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} I_{\{0,1,\ldots\}^n}(y_1, \ldots, y_n)$$

where $\{0, 1, ...\}^n$ denotes the Cartesian product n times over the set $\{0, 1, ...\}$.

The parameter θ is restricted to the space $\Theta = (0, \infty)$.

Distribución previa para θ

- La distribución previa del parámetro θ dada por

$$p(\theta \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} I_{(0,\infty)}(\theta).$$

- La distribución posterior del parámetro θ está dada por

$$\theta \mid Y \sim Gamma\left(\sum_{i=1}^{n} y_i + \alpha, n + \beta\right)$$

Prior Distribution for θ

- The prior distribution for the parameter θ is given by

$$p(\theta \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} I_{(0,\infty)}(\theta).$$

- The posterior distribution for the parameter θ is given by

Estimation Process in **STAN**

Let Y be the count of surveyed people living below the poverty line, expressed as a rate of (X) per 100 inhabitants, by administrative division of the country.

```
dataPois <- readRDS("www/00_Intro_bayes/Poisson/dataPoisson.rds")
```

Table 1: People Count

dam2	n
05002	2
05031	1
05034	1
05045	2
05079	1
05088	6
05093	1
05120	2
05129	1
05142	1

Histogram with People Count

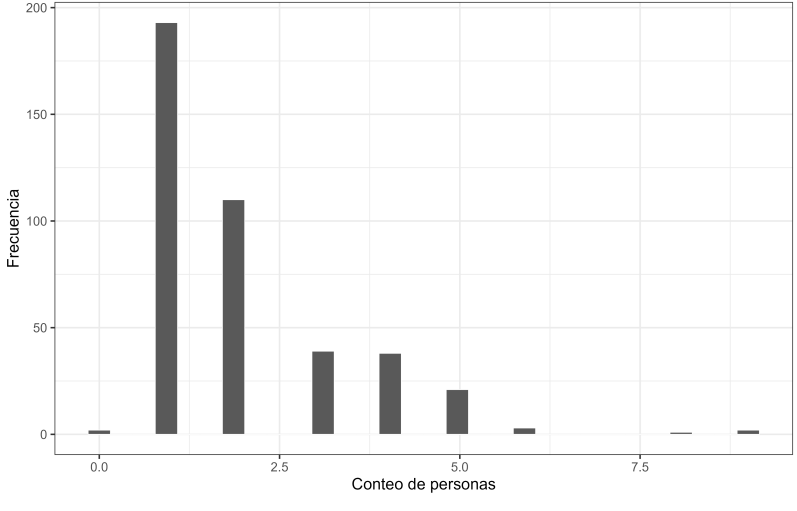


Figure 2: People count by administrative division

Model Written in STAN Code

```
data {
  int<lower=0> n;      // Number of geographic areas
  int<lower=0> y[n];   // Counts per area
  real<lower=0> alpha;
  real<lower=0> beta;
}
parameters {
  real<lower=0> theta;
}
model {
  y ~ poisson(theta);
  theta ~ gamma(alpha, beta);
}
generated quantities {
    real ypred[n];                    // Vector of length n
    for(ii in 1:n){
    ypred[ii] = poisson_rng(theta);
    }
}
```

Preparing Data for STAN Code

- Organizing data for STAN

```
sample_data <- list(n = nrow(dataPois), y = dataPois$n,
                    alpha = 0.001, beta = 0.001)
```

- Running the STAN code

```
stan_pois <- "www/00_Intro_bayes/Poisson/03_Poisson.stan"
model_poisson <-
  stan(
    file = stan_pois, data = sample_data,
    warmup = 500,
    iter = 1000,
    verbose = FALSE,   cores = 4
  )
saveRDS(model_poisson,
          "www/00_Intro_bayes/Poisson/model_poisson.rds")
```

Results of the Estimation of Parameter θ

```
tabla_posi <- summary(model_poisson,
                      pars =c("theta"))$summary
tabla_posi%>%tba()
```

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
theta	2.03	0.0023	0.0685	1.901	1.982	2.032	2.077	2.166	851.6	1.002

Convergence of Chains for Parameter θ

```
posterior_theta <- as.array(model_poisson, pars = "theta")
p1 <- (mcmc_dens_chains(posterior_theta) +
    mcmc_areas(posterior_theta) ) / mcmc_trace(posterior_theta)
```

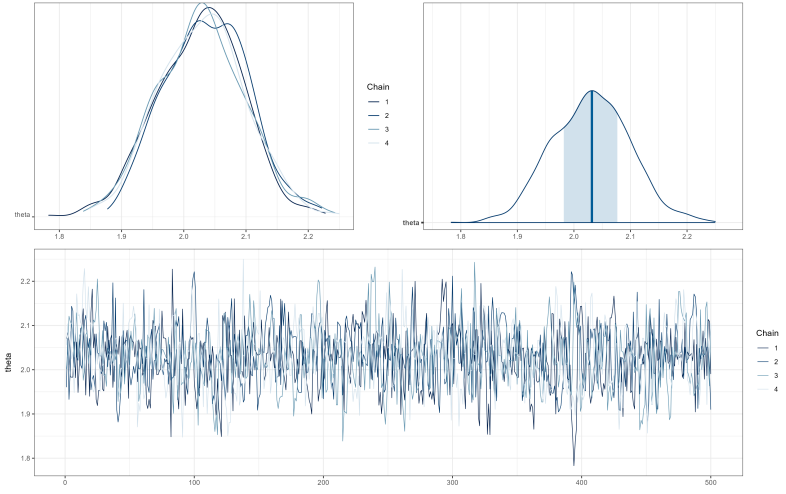


Figure 3: Chains for theta

Posterior Predictive Check

```
y_pred_B<-as.array(model_poisson,pars ="ypred") %>%
  as_draws_matrix()

rowsrandom<-sample(nrow(y_pred_B),100)

y_pred2<-y_pred_B[rowsrandom,]

p1<- ppc_dens_overlay(y =as.numeric(dataPois$n*100), y_pred2*100)
```

Posterior Predictive Check

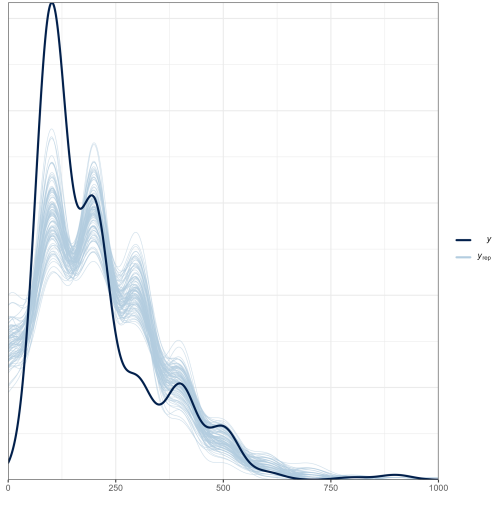


Figure 4: Chains for theta

Unit Model: Normal with Unknown Mean and Variance

- In the normal model, a set of independent and identically distributed variables $Y_1, \cdots, Y_n \sim N(\theta, \sigma^2)$ is considered.
- When both the mean and the variance of the distribution are unknown, different approaches are proposed to assign prior distributions to θ and σ^2 based on the problem context.

Prior Distributions for θ and σ^2

Three possible assumptions about prior distributions for θ and σ^2 are described, considering independence and informativeness.

- Assume that the prior distribution $p(\theta)$ is independent of the prior distribution $p(\sigma^2)$ and that both distributions are informative.
- Assume that the prior distribution $p(\theta)$ is independent of the prior distribution $p(\sigma^2)$ and that both distributions are non-informative.
- Assume that the prior distribution for θ depends on σ^2 and write it as $p(\theta \mid \sigma^2)$, while the prior distribution for σ^2 does not depend on θ and can be written as $p(\sigma^2)$.

The prior distribution is set for the parameter θ as $\theta \sim Normal(0, 10000)$ and for the parameter σ^2 as $\sigma^2 \sim IG(0.0001, 0.0001)$.

Definition of the Normal Model

- The goal of the model is to estimate the average income of people, represented as $\bar{Y}_d = \frac{\sum_{U_d} y_{di}}{N_d}$.
- A way to estimate $\bar{Y}$ using $\hat{y}_{di}$, which is the expected value of y_{di} under a probability measure induced by the model, is shown.
- Finally, the estimate of $\hat{\bar{Y}}_d = \frac{\sum_{U_d} \hat{y}_{di}}{N_d}$ is presented.

Estimation Process

- To estimate the average income of people, i.e.,

$$\bar{Y}_d = \frac{\sum_{U_d} y_{di}}{N_d}$$

where y_{di} is the income of each person. Note that

$$\bar{Y}_d = \frac{\sum_{s_d} y_{di} + \sum_{s_d^c} y_{di}}{N_d}$$

Now, the estimator of $\bar{Y}$ is given by:

$$\hat{\bar{Y}}_d = \frac{\sum_{s_d} y_{di} + \sum_{s_d^c} \hat{y}_{di}}{N_d}$$

Estimation Process

Now, it is possible to assume that $\hat{y}_{di}$ is the conditional expectation given the modeling, that is

$$\hat{y}_{di} = E_{\mathcal{M}}\left(y_{di} \mid x_d, \beta\right)$$

,

where $\mathcal{M}$ refers to the probability measure induced by the modeling. Finally, it is found that

$$\hat{\bar{Y}}_d = \frac{\sum_{U_d} \hat{y}_{di}}{N_d}$$

Estimation Process in **STAN**

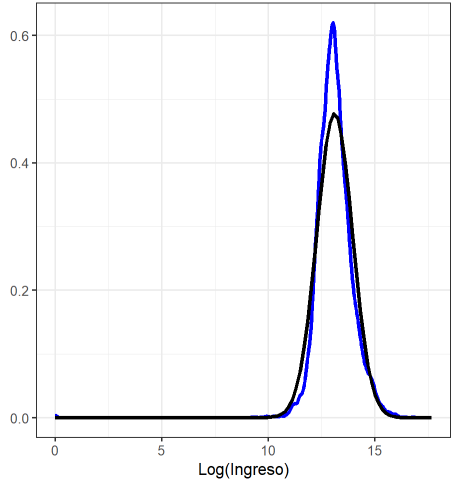
Let Y be the logarithm of income for an administrative division of the country.

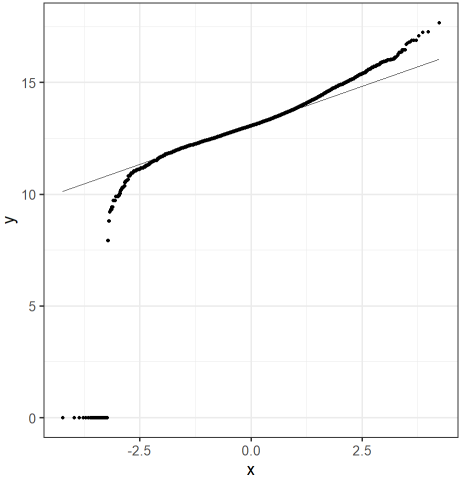
```
dataNormal <- readRDS("www/00_Intro_bayes/Normal/01_dataNormal.rds")
tba(dataNormal %>% head(10), cap = "Logarithm of income" )
```

Table 2: Logarithm of income

dam_ee	logIngreso
08	14.37
08	14.37
08	14.37
08	12.02
08	12.02
08	12.02
08	14.05
08	14.05
08	14.05
08	14.05

Graphical Analysis of Logarithm of Income





Model Written in STAN Code

```
data {
  int<lower=0> n;
  real y[n];
}
parameters {
  real sigma;
  real theta;
}
transformed parameters {
  real sigma2;
  sigma2 = pow(sigma, 2);
}
```

```
model {
  y ~ normal(theta, sigma);
  theta ~ normal(0, 1000);
  sigma2 ~ inv_gamma(0.001, 0.001);
}
generated quantities {
    real ypred[n];
    for(kk in 1:n){
    ypred[kk] = normal_rng(theta,sigma);
}
}
```

Preparing Data for the STAN Code

- Organizing data for STAN

```
sample_data <- list(n = nrow(dataNormal),
                    y = dataNormal$logIngreso)
```

- Running STAN from R using the **rstan** library

```
NormalMeanVar  <- "www/00_Intro_bayes/Normal/03_NormalMeanVar.stan"
model_NormalMedia <- stan(
  file = NormalMeanVar,
  data = sample_data,
  warmup = 500,
  iter = 1000,
  verbose = FALSE, cores = 4
)
saveRDS(model_NormalMedia,
        "www/00_Intro_bayes/Normal/model_NormalMedia2.rds")
```

Results of the Estimation of the Parameters θ and σ^2 are:

```
tabla_Nor2 <- summary(model_NormalMedia,
        pars = c("theta", "sigma2", "sigma"))$summary
```

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff
theta	13.1147	1e-04	0.0039	13.1068	13.1120	13.1148	13.1172	13.1224	1221
sigma2	0.6986	1e-04	0.0047	0.6894	0.6955	0.6986	0.7016	0.7080	1931
sigma	0.8358	1e-04	0.0028	0.8303	0.8340	0.8358	0.8376	0.8415	1931

Convergence of Chains for Parameter θ

```
posterior_theta <- as.array(model_NormalMedia, pars = "theta")
(mcmc_dens_chains(posterior_theta) +
    mcmc_areas(posterior_theta) ) /
  mcmc_trace(posterior_theta)
```

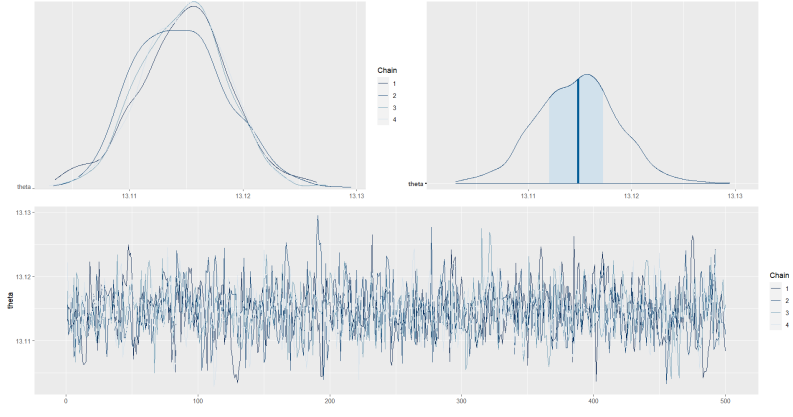


Figure 5: Chains for theta

Convergence of Chains for Parameter σ^2

```
posterior_sigma2 <- as.array(model_NormalMedia, pars = "sigma2")
(mcmc_dens_chains(posterior_sigma2) +
    mcmc_areas(posterior_sigma2) ) /
  mcmc_trace(posterior_sigma2)
```

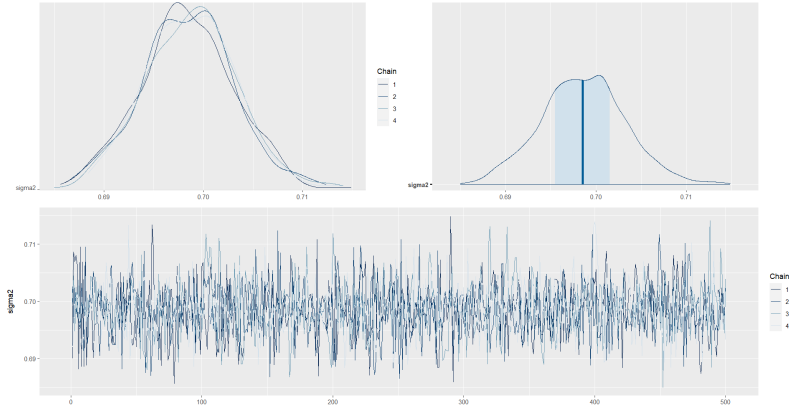


Figure 6: Chains for sigma2

Posterior Predictive Check of Income

```
y_pred_B <- as.array(model_NormalMedia, pars = "ypred") %>%
  as_draws_matrix()

rowsrandom <- sample(nrow(y_pred_B), 100)

y_pred2 <- y_pred_B[rowsrandom,]

ppc_dens_overlay(
  y = as.numeric(exp(dataNormal$logIngreso) - 1), y_pred2) +
  xlim(0, 5000000)
```

Posterior Predictive Check of Income

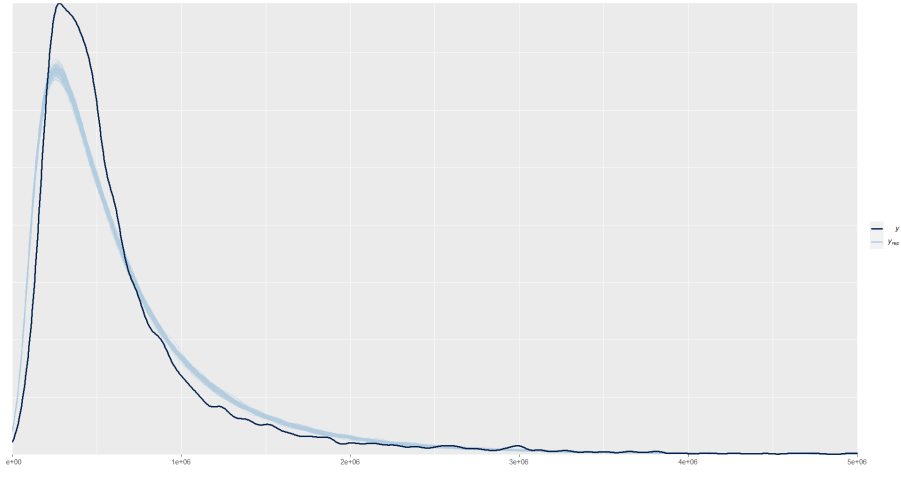


Figure 7: PPC for income

Linear Models.

Linear regression is the basic technique in econometric analysis. Through this technique, we aim to determine linear dependence relationships between a dependent variable, or endogenous variable, and one or more explanatory variables, or exogenous variables.

Bayesian Linear Models

First, note that the particular interest lies in the distribution of the vector of n random variables $Y = (Y_1, \dots, Y_n)'$ conditioned on the matrix of auxiliary variables X and indexed by the vector of parameters of interest $\beta = (\beta_1, \dots, \beta_q)'$ given by $p(Y \mid \beta, X$.

The basic and classical model assumes that the likelihood for the variables of interest is:

$$Y \mid \theta, \sigma^2, X \sim \text{Normal}_n(X\beta, \sigma^2 I_n)$$

where I_n denotes the identity matrix of order $n \times n$. Of course, the normal model is not the only one that can be postulated as the likelihood for the data.

Independent Parameters

Assuming that the parameters are independent a priori, meaning that the joint prior distribution is given by:

$$p(\beta, \sigma^2) = p(\beta)p(\sigma^2)$$

Naturally, the prior distribution for the parameter vector β is normal. However, this time, the variance-covariance matrix will not depend on the other parameter σ^2. So, you have:

$$\beta \sim \text{Normal}_q(b, B)$$

Similarly, the parameter σ^2 does not depend on β, and you can assign it the following prior distribution:

$$\sigma^2 \sim Inverse - Gamma\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right)$$