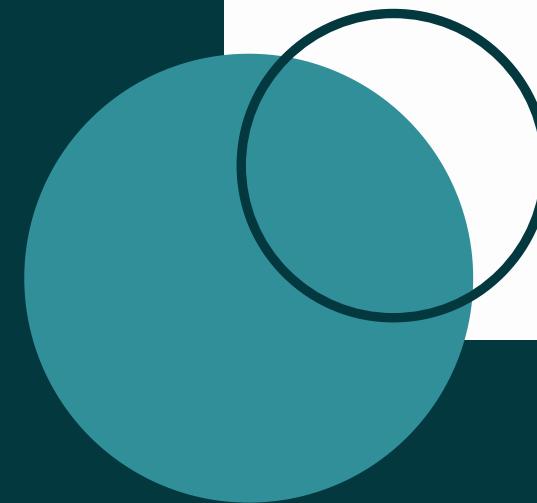


EQUIPO 10

REGRESIÓN BAYESIANA



REGRESIÓN BAYESIANA

La regresión lineal Bayesiana univariada es un enfoque de Regresión lineal donde el análisis estadístico se realiza dentro del contexto de la inferencia Bayesiana.

Objetivo

- Determinar la distribución posterior de los parámetros del modelo.
- Determinar la verosimilitud de los datos.
- Aplicar el teorema de Bayes para actualizar la distribución a priori en forma de distribución a posteriori.

TEOREMA DE BAYES

La probabilidad de un modelo dado por los nuevos datos es igual a la probabilidad de los nuevos datos dado un modelo por el modelo actual dividido por la probabilidad de los nuevos datos.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

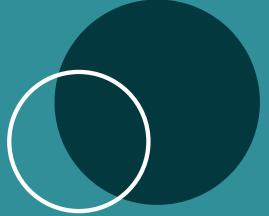
↑
Posterior
↓
 $P(A|B)$

Likelihood
↓
 $P(B|A)$

Prior
↓
 $P(A)$

↑
Evidence
↑

PRINCIPALES FACTORES



$$Posterior = \frac{Likelihood * Prior}{Normalization}$$

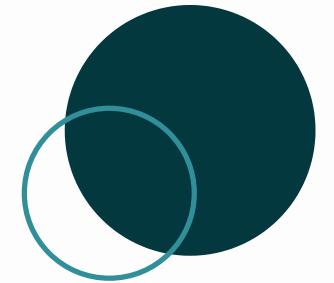
Aquí podemos observar las dos principales ventajas de la regresión lineal en la formula del teorema de Bayes que son:

Prioridades (prior): Suposición de cuáles deberían ser los parámetros del modelo.

Posterior: El resultado de realizar una regresión lineal bayesiana.



DIFERENCIAS



REGRESIÓN LINEAL CLÁSICA

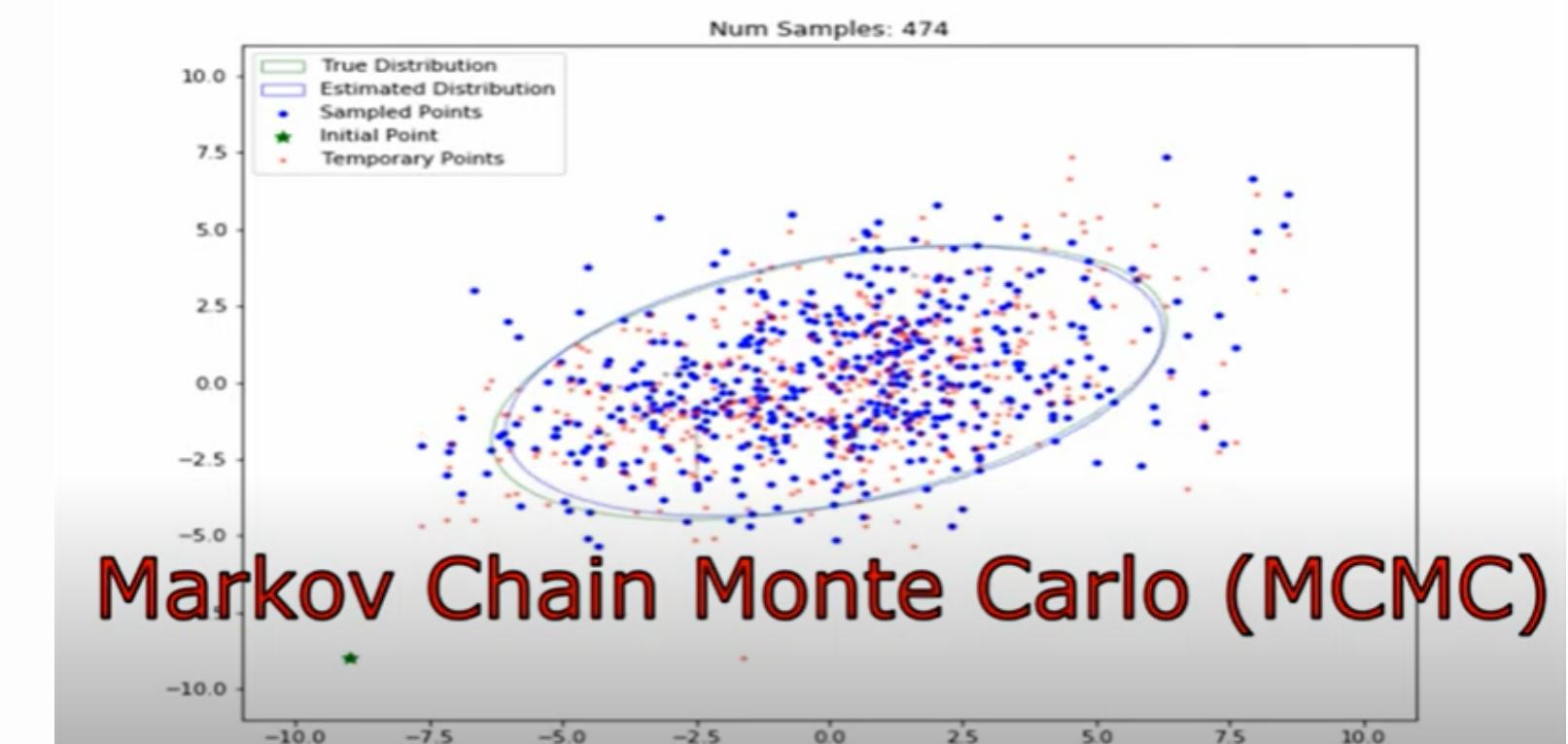
- Trata de encontrar un modelo que encaje con los datos
- Utiliza el método de los mínimos cuadrados
- Toma en cuenta los datos que tenemos

REGRESIÓN BAYESIANA

- Trata de encontrar una distribución que encaje con los datos
- El metodo mas común utilizado es el MCMC
- Toma en cuenta los datos que tenemos y tus propias creencias

CADENA DE MARKOV MONTE CARLO (MCMC)

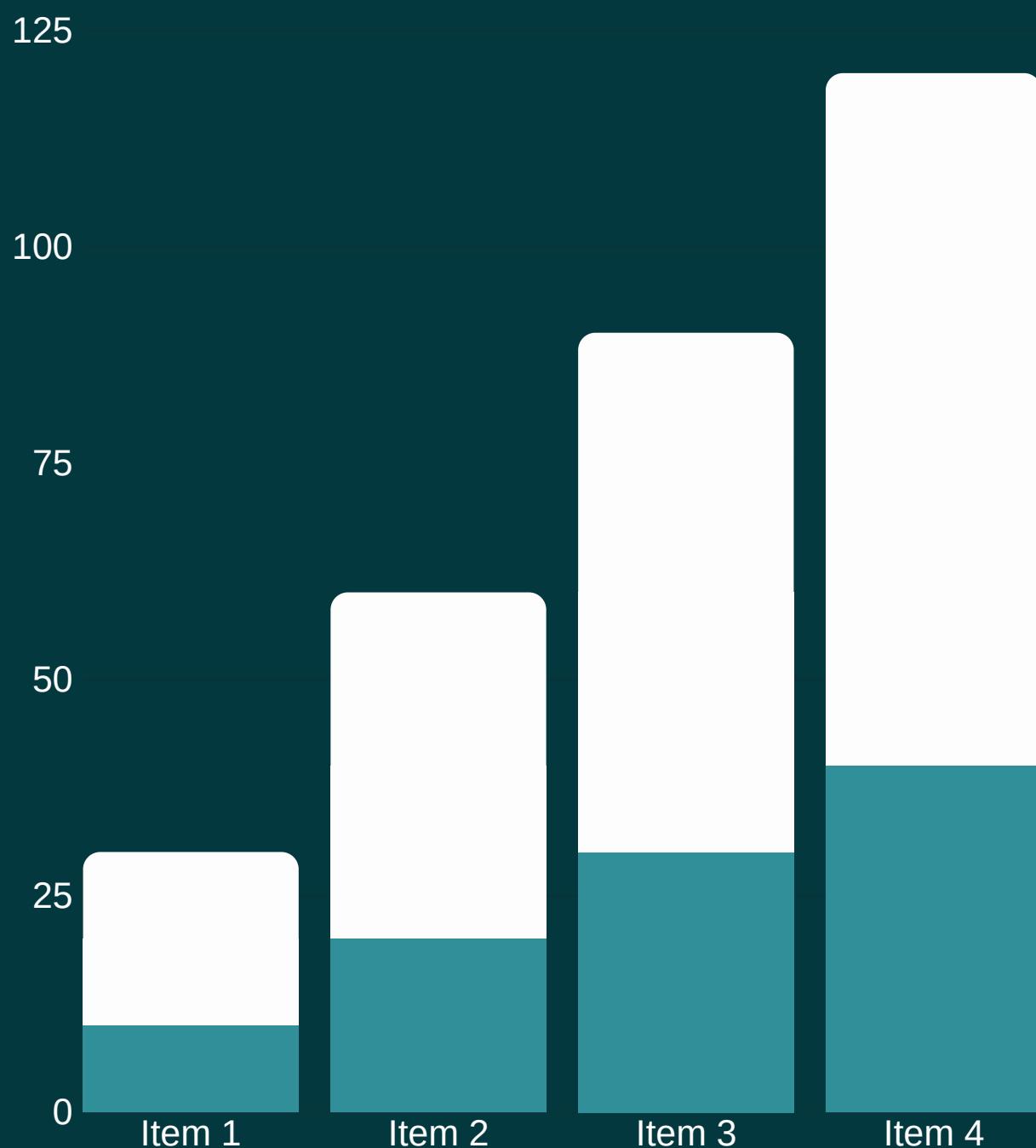
La regresión Bayesiana usa muchos tipos de técnicas de muestreo, la aproximación más popular es el "Muestreo de Gibbs" también conocido como Cadena de Markov de Monte Carlo



Donde se obtiene una secuencia de observaciones que son aproximadas apartir de cierta distribución de probabilidad

CONDICIONALES COMPLETOS

Usamos las formulas de condicionales completos para simular distribuciones multivariadas de nuestros parametros. Esencialmente el condicional completo es la distribución condicional sobre todas las demás variables

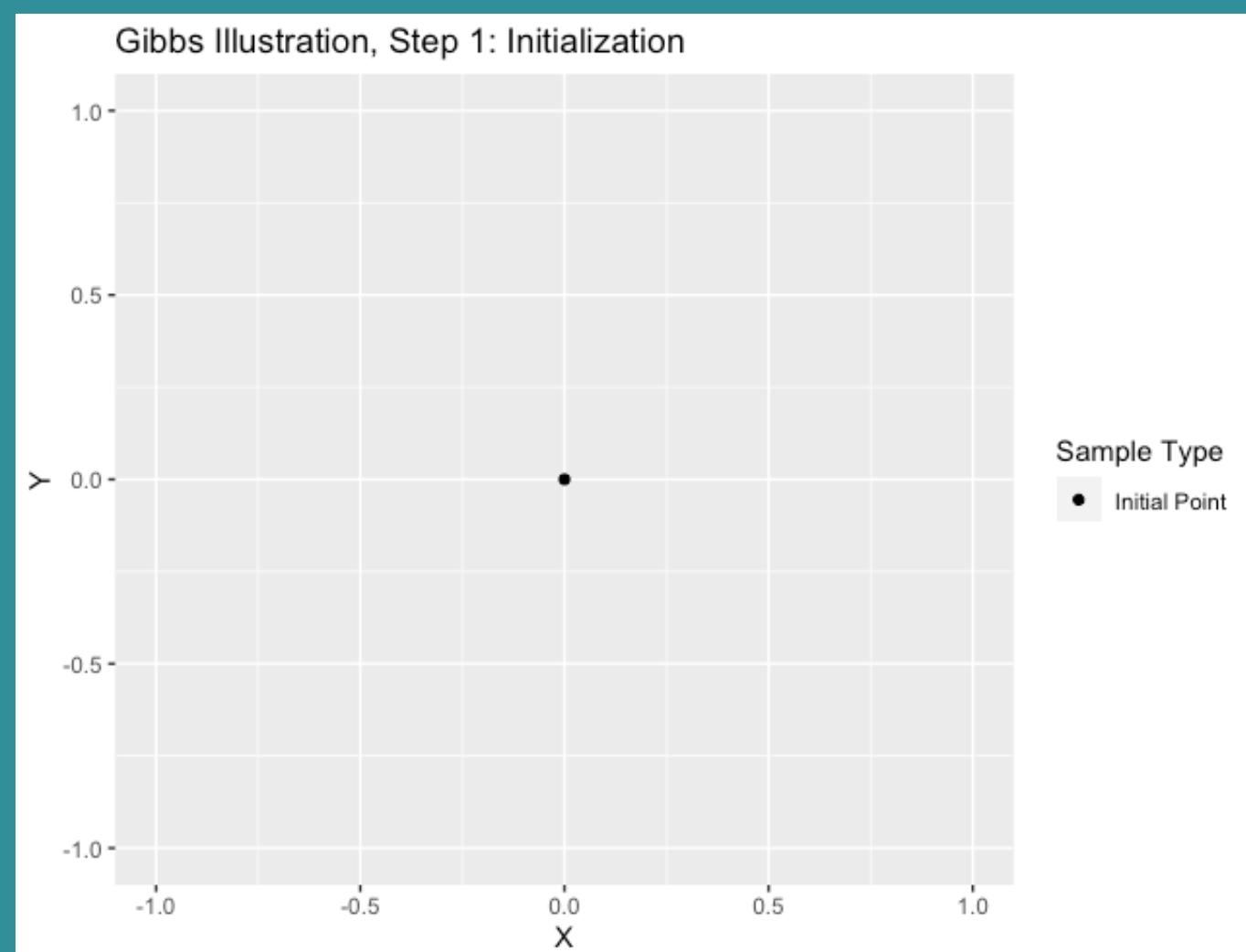


$$X \mid (Y = y) \sim N(\rho y, 1 - \rho^2)$$

$$Y \mid (X = x) \sim N(\rho x, 1 - \rho^2)$$

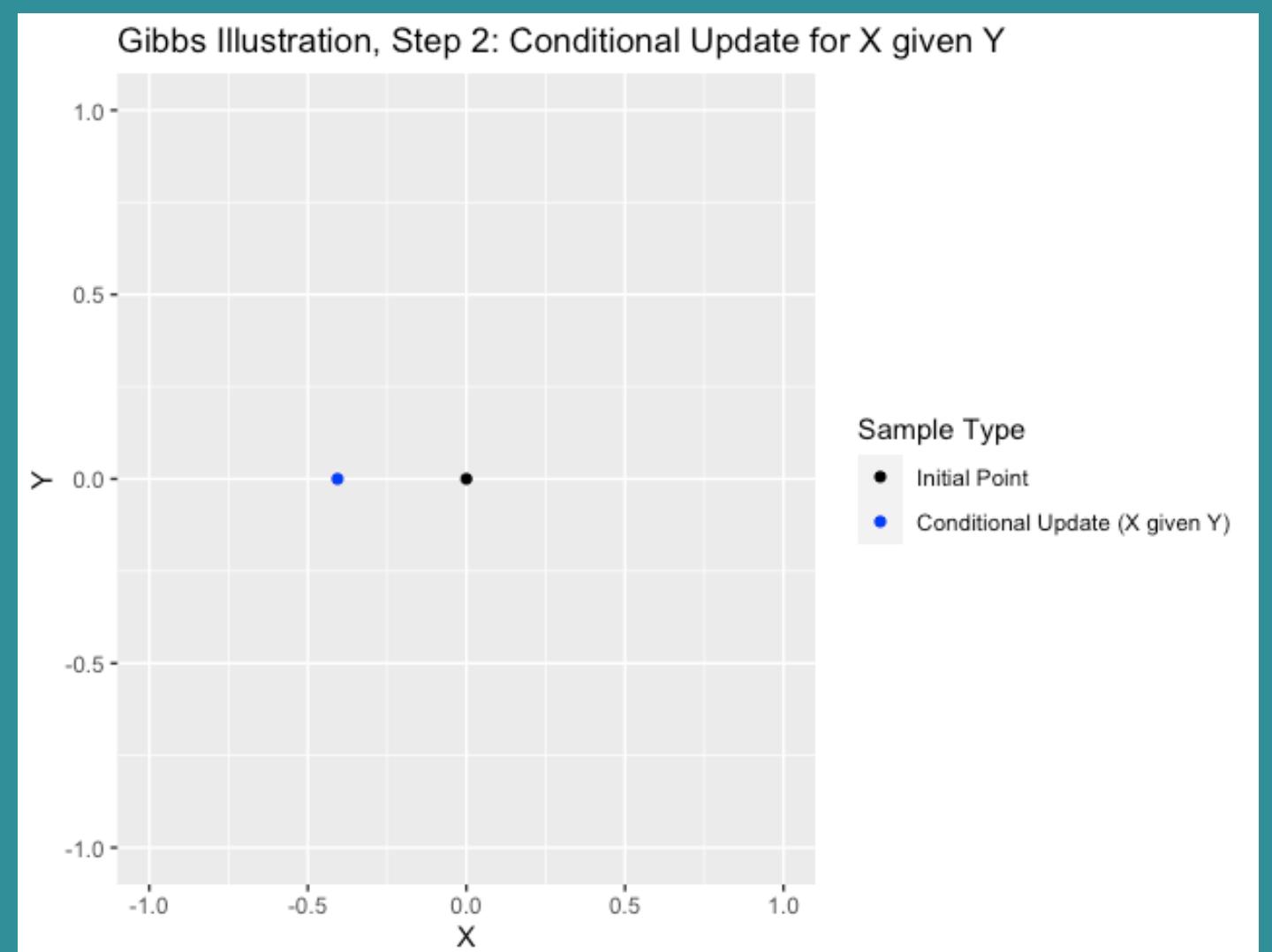
¿Cómo funciona?

Paso 1: inicialización
 inicializar (x_0, y_0) en $(0,0)$ y establecer el contador de iteración t en 0.



Paso 2: Actualización condicional de X dado Y

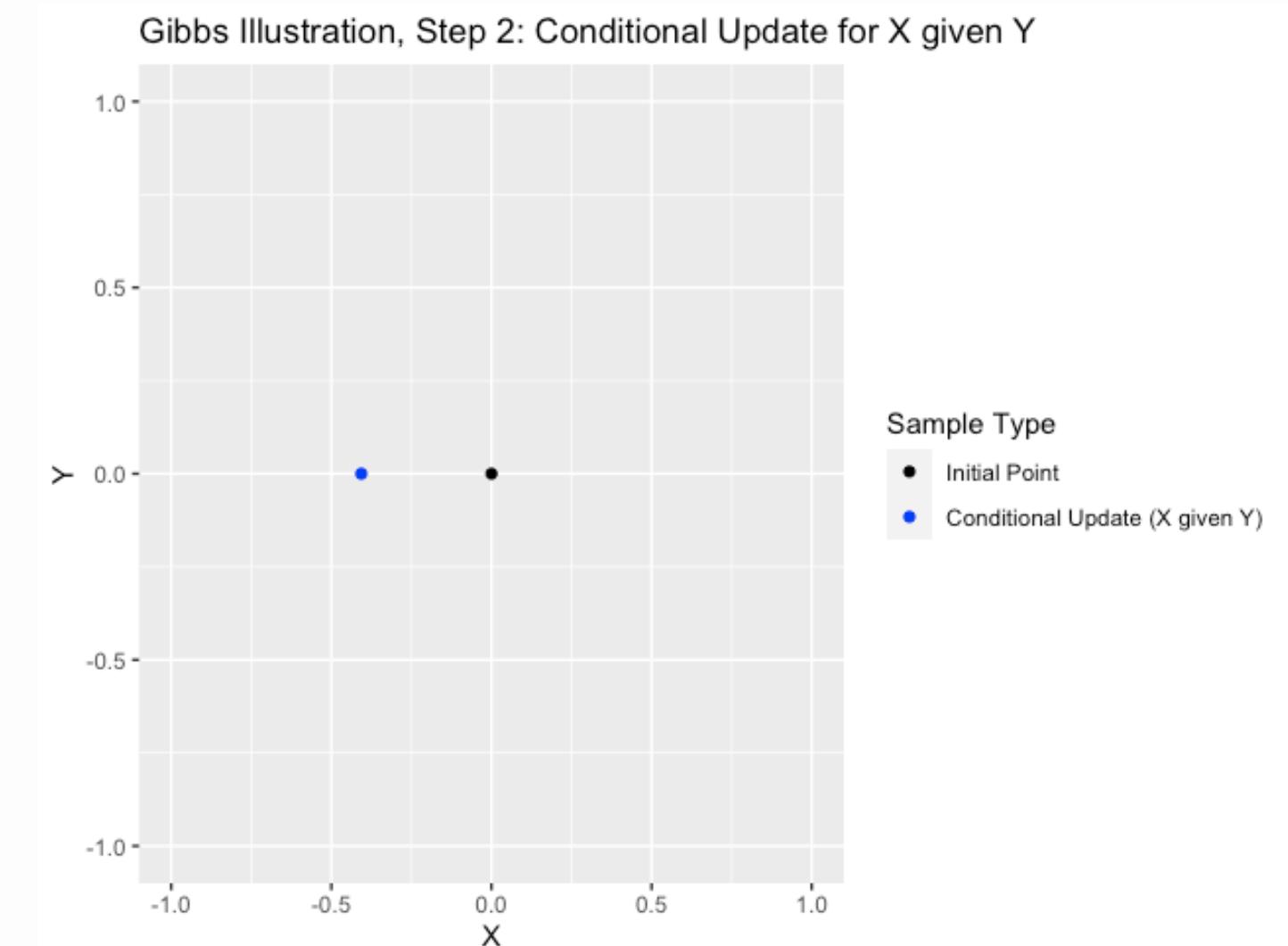
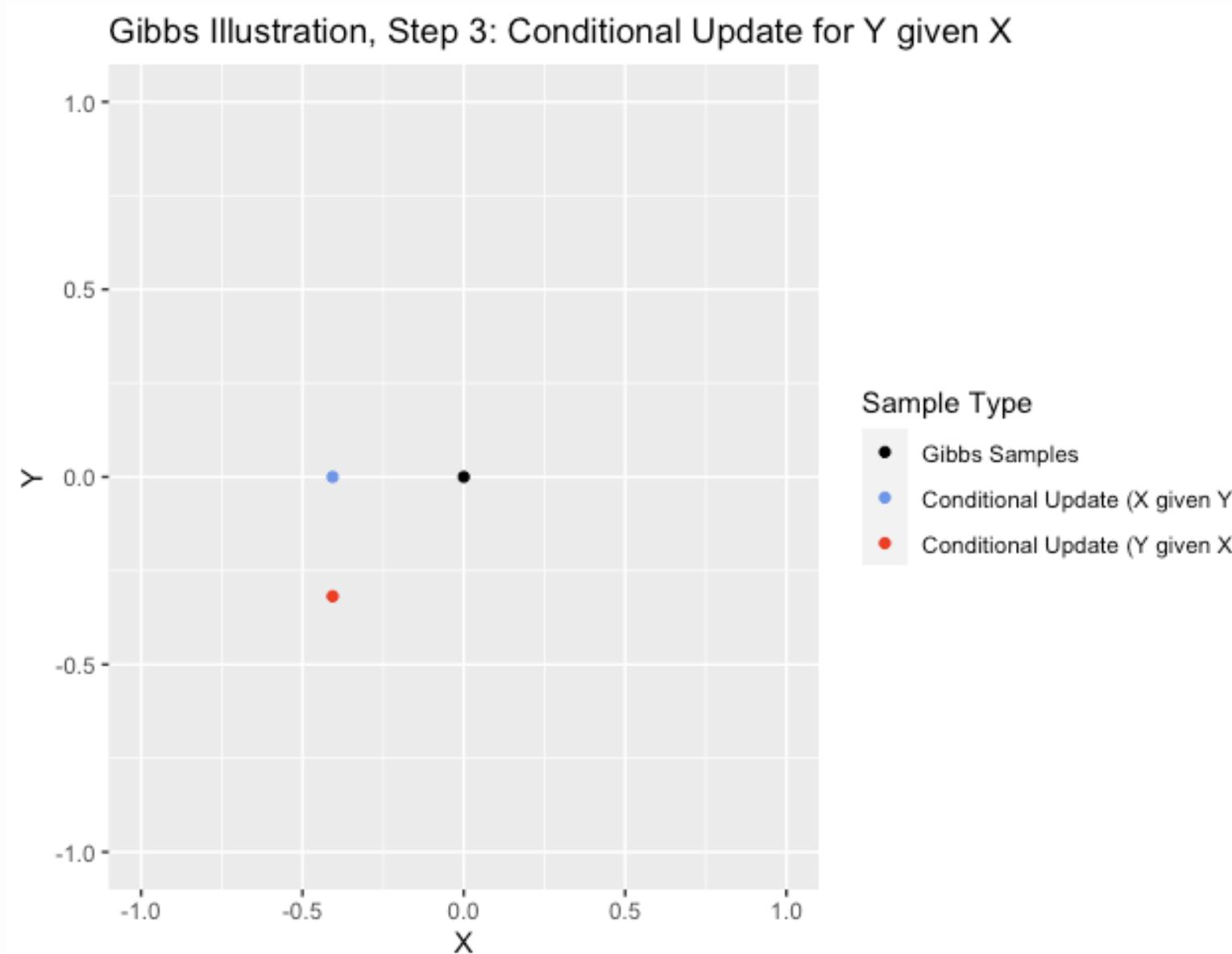
$$X_1 \mid (Y_0 = 0) \sim N(0 \cdot \rho, 1 - \rho^2)$$



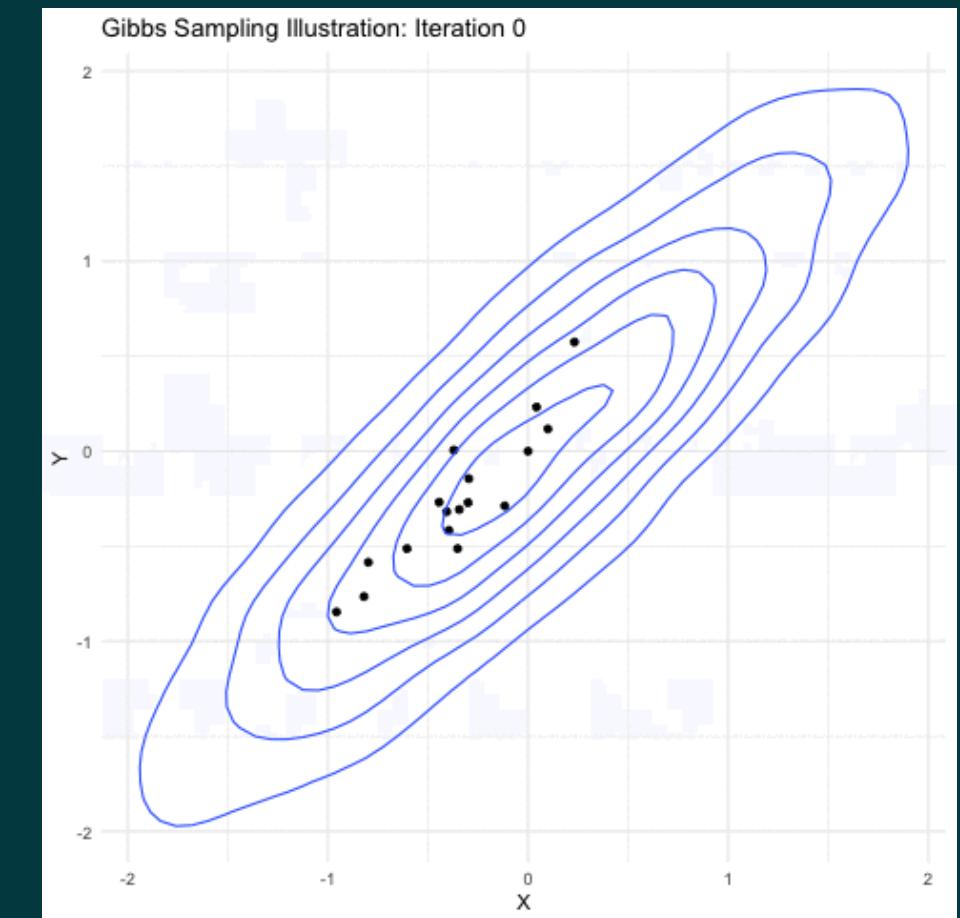
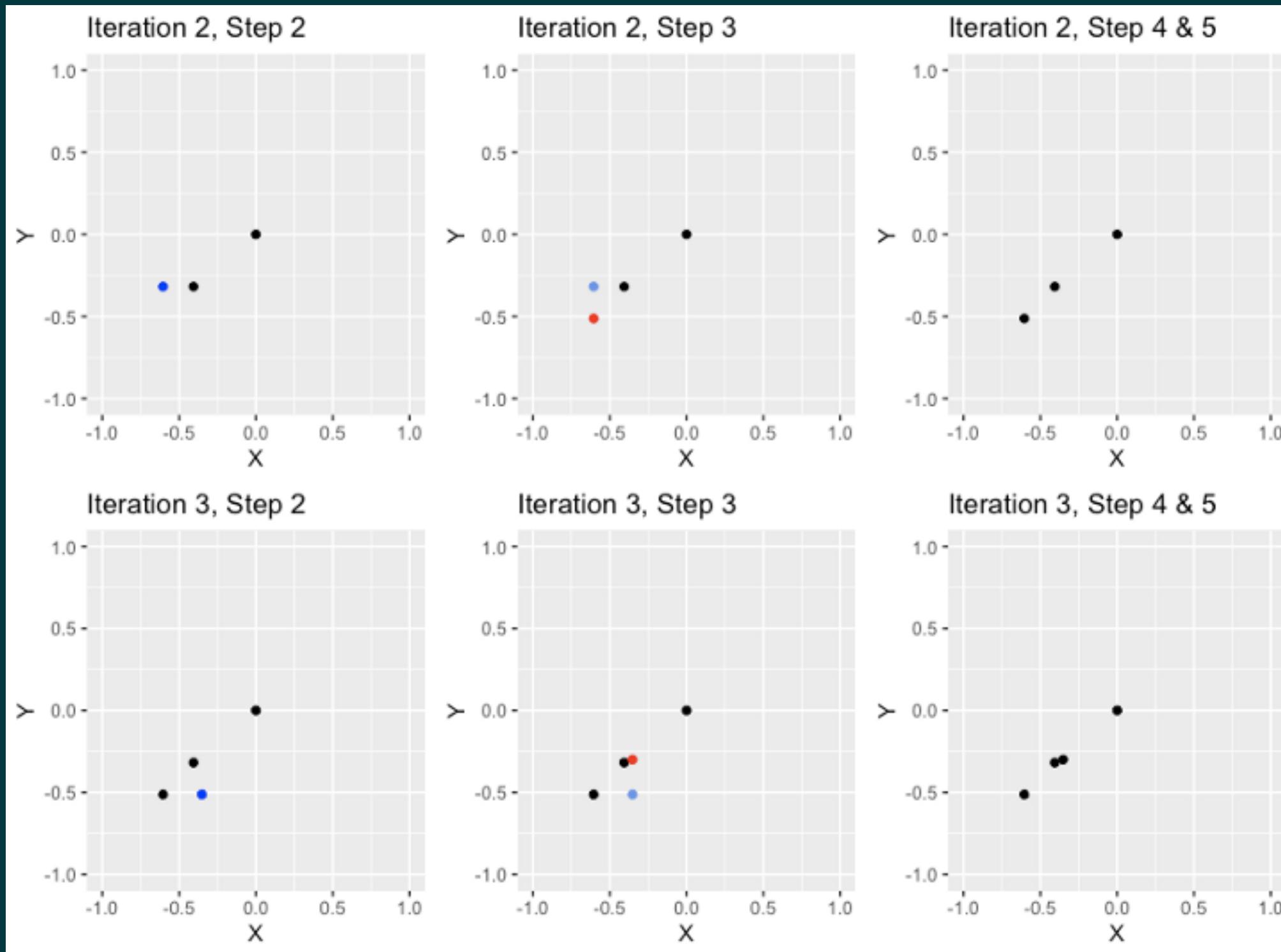
Paso 3: Actualización condicional de Y dado X

$$Y_1 \mid (X_1 = -0.4) \sim N(-0.4 \cdot \rho, 1 - \rho^2)$$

Paso 4 y 5: incremente el contador de iteraciones y regrese al paso 2.

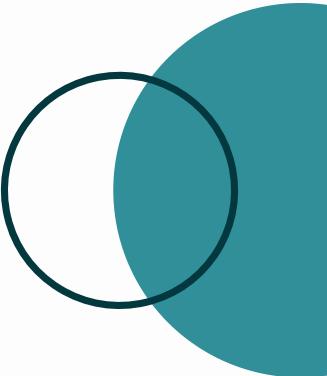


MÁS ITERACIONES



VENTAJAS

- No necesitas toneladas de datos para encontrar un buen ajuste
- Es muy flexible con buena información
- Resuelve problemas que por lo general no se pueden resolver usando las aproximaciones a las frecuencias
- Similar a la intuición humana





DATOS CURIOSOS

-

NATE SILVER

Utiliza muchas ideologías bayesianas para elaborar los resultados de las elecciones de los cuales siempre resultan correctos

- Si tienes un número infinito de datos los parametros convergen a los valores vistos en la regresión lineal clasica

Ejemplo 1

Base de datos Iris

```
#Regresion Bayesiana  
  
#Ejemplo 1:  
  
install.packages("MCMCpack")  
library(MCMCpack)  
  
tabla <- data.frame(iris$Sepal.Length,iris$Sepal.Width)  
bayes <- MCMCregress(tabla$iris.Sepal.Width~tabla$iris.Sepal.Length,data = tabla)
```

Instalamos la libreria MCMCpack que es la que trae las funciones para poder realizar la cadena de markov de Monte Carlo

Así como tambien guardamos la tabla y hacemos la regresión con la fucion MCMCregress

```
summary(bayes)  
plot(bayes)
```

Pedimos el resumen del modelo de la regresión Bayesiana y graficamos

Resumen del modelo

```
> tabla <- data.frame(iris$Sepal.Length,iris$Sepal.Width)
> bayes <- MCMCregress(tabla$iris.Sepal.Width~tabla$iris.Sepal.Length,data = tabla)
>
> summary(bayes)
```

Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

| | Mean | SD | Naive SE | Time-series SE |
|--------------------------|----------|---------|-----------|----------------|
| (Intercept) | 3.42046 | 0.25526 | 0.0025526 | 0.0025526 |
| tabla\$iris.Sepal.Length | -0.06214 | 0.04333 | 0.0004333 | 0.0004333 |
| sigma2 | 0.19131 | 0.02265 | 0.0002265 | 0.0002312 |

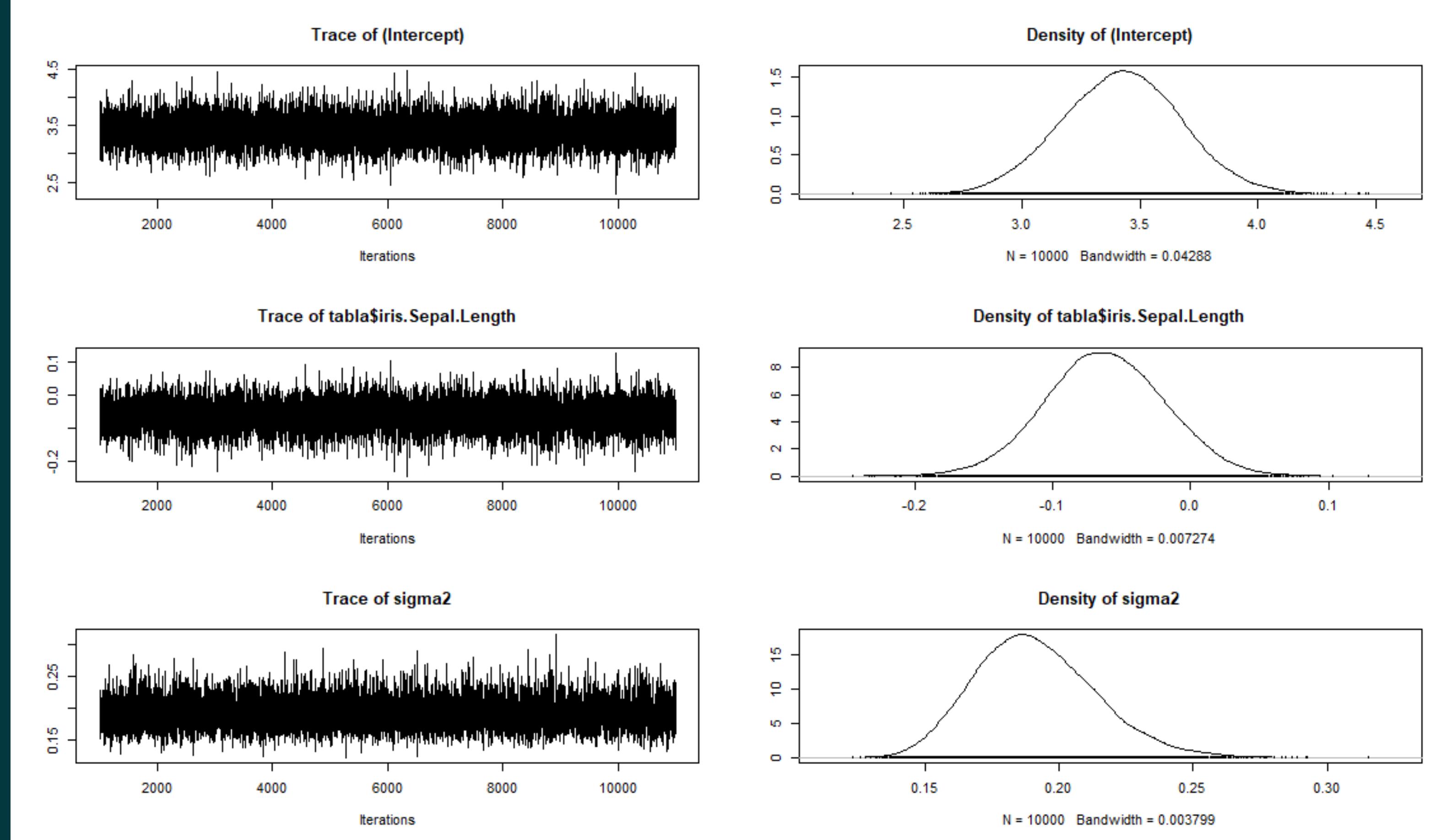
La regresión Bayesiana nos dio de resultado los parametros de media y desviación

2. Quantiles for each variable:

| | 2.5% | 25% | 50% | 75% | 97.5% |
|--------------------------|---------|----------|----------|----------|---------|
| (Intercept) | 2.9244 | 3.24900 | 3.42280 | 3.59124 | 3.92360 |
| tabla\$iris.Sepal.Length | -0.1478 | -0.09094 | -0.06247 | -0.03292 | 0.02159 |
| sigma2 | 0.1519 | 0.17530 | 0.18950 | 0.20560 | 0.24037 |

```
>  
> plot(bayes)
```

Aqui tenemos los cuantiles de cada variable



del lado izquierdo podemos ver todas las observaciones y como se comportan, del lado derecho su respectiva gráfica

Regresión lineal clásica modelo

```
> bayes_clasica <- lm(tabla$iris.Sepal.Width~tabla$iris.Sepal.Length,data = tabla)
> summary(bayes_clasica)
```

Call:

```
lm(formula = tabla$iris.Sepal.Width ~ tabla$iris.Sepal.Length,
  data = tabla)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|---------|---------|--------|--------|
| -1.1095 | -0.2454 | -0.0167 | 0.2763 | 1.3338 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|--------------------------|----------|------------|---------|------------|
| (Intercept) | 3.41895 | 0.25356 | 13.48 | <2e-16 *** |
| tabla\$iris.Sepal.Length | -0.06188 | 0.04297 | -1.44 | 0.152 |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4343 on 148 degrees of freedom

Multiple R-squared: 0.01382, Adjusted R-squared: 0.007159

F-statistic: 2.074 on 1 and 148 DF, p-value: 0.1519

Ejemplo 2

```
> getSymbols("AUDNZD=X",src="yahoo",from = "2017-02-18")
[1] "AUDNZD=X"
Warning message:
AUDNZD=X contains missing values. Some functions will not work if objects contain missing values in the middle of the series. Consider using na.omit(), na.approx(), na.fill(), etc to remove or replace them.
> getSymbols("AUDUSD=X",src="yahoo",from = "2017-02-18")
[1] "AUDUSD=X"
Warning message:
AUDUSD=X contains missing values. Some functions will not work if objects contain missing values in the middle of the series. Consider using na.omit(), na.approx(), na.fill(), etc to remove or replace them.
> AUDNZD= `AUDNZD=X` [,4]
> AUDUSD= `AUDUSD=X` [,4]
> X=as.numeric(AUDNZD)
> Y=as.numeric(AUDUSD)
```

Se realiza otro ejemplo con otras variables

```
> bayes <- MCMCregress(Y~X)
> summary(bayes)
```

Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

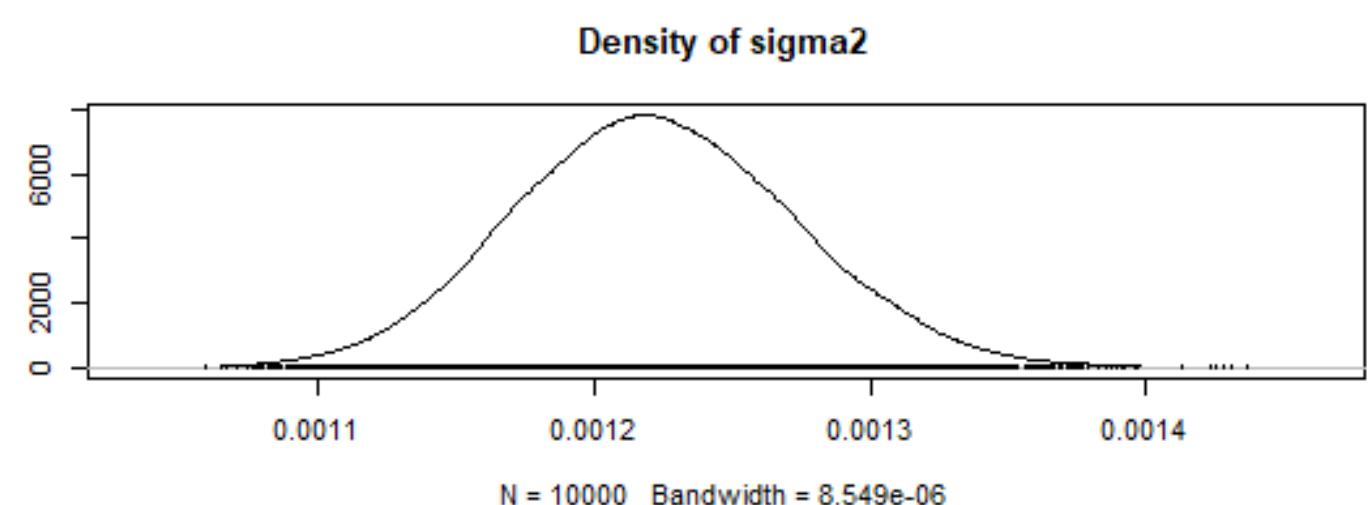
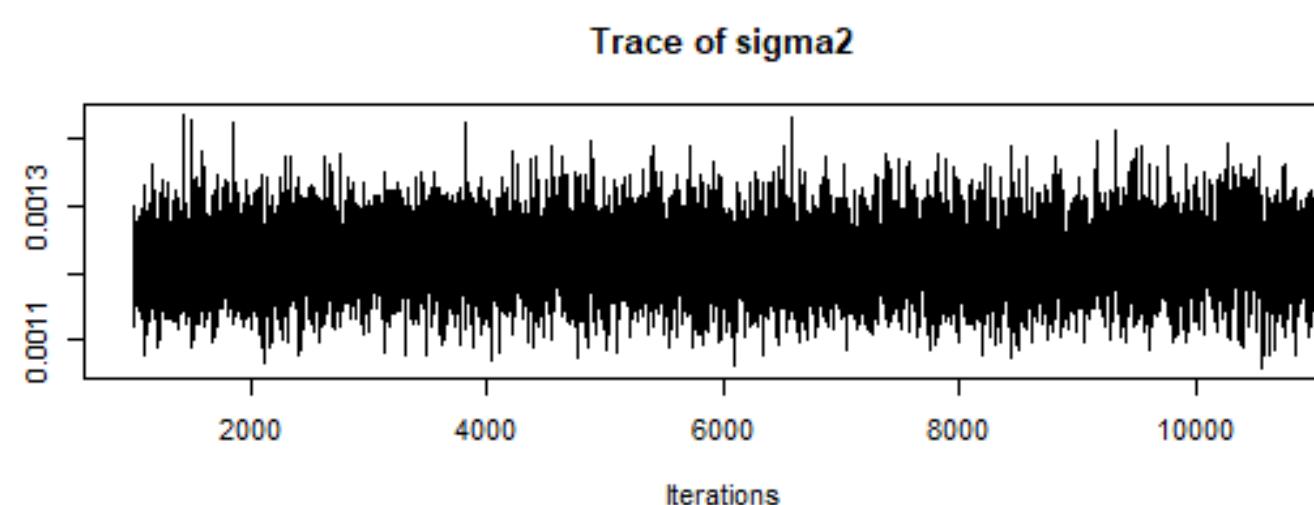
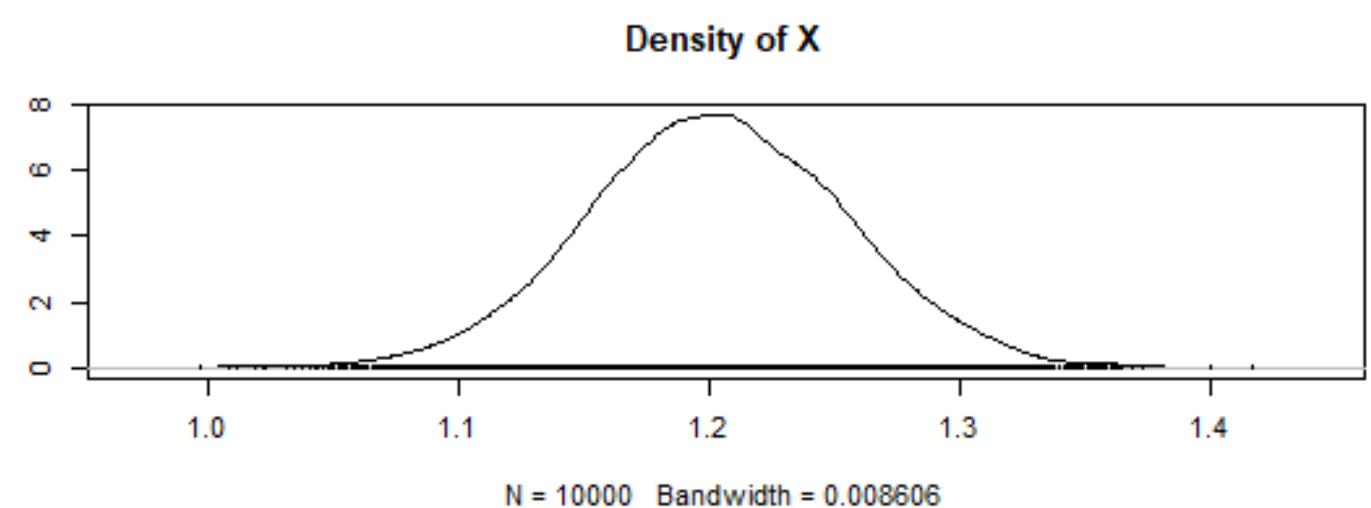
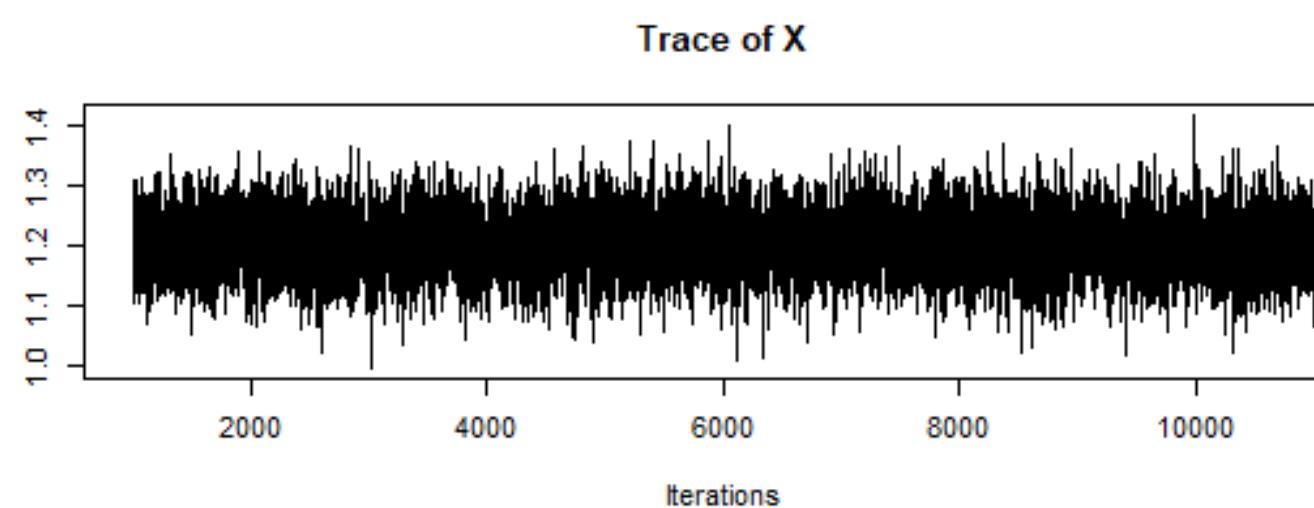
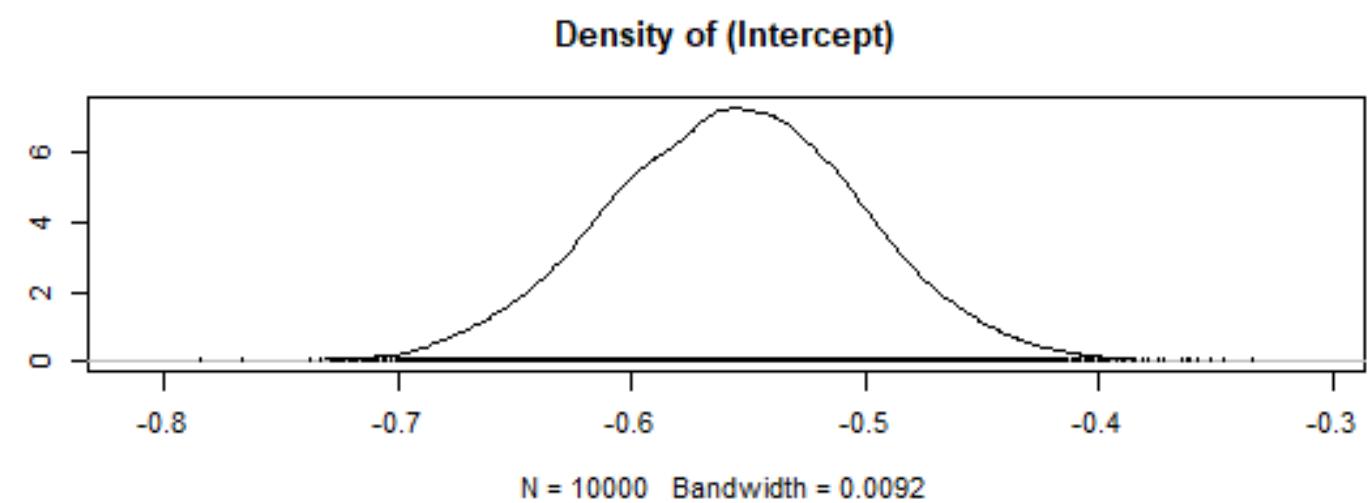
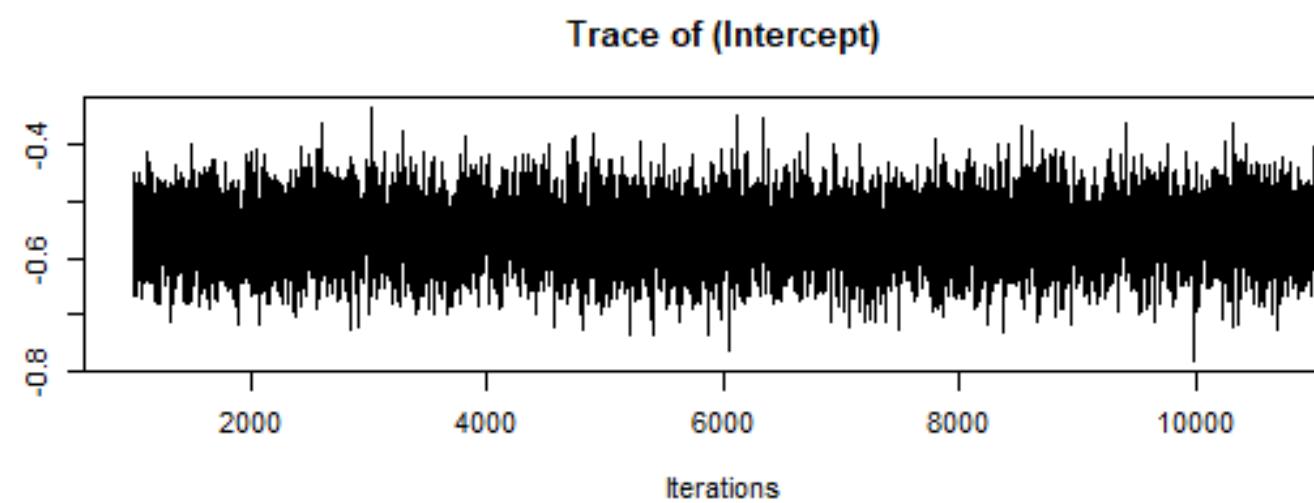
| | Mean | SD | Naive SE | Time-series SE |
|-------------|-----------|-----------|-----------|----------------|
| (Intercept) | -0.555917 | 5.476e-02 | 5.476e-04 | 5.476e-04 |
| X | 1.203355 | 5.122e-02 | 5.122e-04 | 5.122e-04 |
| sigma2 | 0.001223 | 5.089e-05 | 5.089e-07 | 5.089e-07 |

Obtenemos el summary para mejor comprensión en los datos y posteriormente hacemos el grafico

2. Quantiles for each variable:

| | 2.5% | 25% | 50% | 75% | 97.5% |
|-------------|-----------|-----------|-----------|-----------|-----------|
| (Intercept) | -0.663358 | -0.592856 | -0.555545 | -0.519027 | -0.448424 |
| x | 1.102844 | 1.168834 | 1.202908 | 1.237918 | 1.304191 |
| sigma2 | 0.001126 | 0.001188 | 0.001221 | 0.001256 | 0.001325 |

Como se observa, aquí se están utilizando más datos y de igual forma como en el ejemplo anterior, se observa la función de densidad con su mejor parámetro.



```
> bayes_clasica <- lm(Y~X)
> summary(bayes_clasica)
```

Call:
lm(formula = Y ~ X)

Residuals:

| Min | 1Q | Median | 3Q | Max |
|-----------|-----------|----------|----------|----------|
| -0.103909 | -0.025014 | 0.001145 | 0.028785 | 0.076836 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|------------|
| (Intercept) | -0.55605 | 0.05475 | -10.16 | <2e-16 *** |
| X | 1.20348 | 0.05120 | 23.51 | <2e-16 *** |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03492 on 1161 degrees of freedom
(23 observations deleted due to missingness)

Multiple R-squared: 0.3224, Adjusted R-squared: 0.3219

F-statistic: 552.5 on 1 and 1161 DF, p-value: < 2.2e-16

Preguntas

- ¿Cuales son las dos principales ventajas de la regresión lineal que se encunetran en la formula del teorema de Bayes?
- ¿Qué diferencias hay entre la regresión lineal clásica y la regresión bayesiana?
- ¿Cuál es la aproximación más popular que usa la regresión bayesiana?
- ¿Para qué se usan las formulas de condicionales completos?
- ¿Cuáles son las ventajas de la regresión bayesiana?



BIBLIOGRAFÍA

- <https://www.youtube.com/watch?v=s1Pm0oHPGT4>
- <https://www.ibm.com/docs/es/spss-statistics/25.0.0?topic=statistics-bayesian-inference-about-linear-regression-models#:~:text=La%20regresión%20lineal%20Bayesiana%20univariada,y%20definir%20un%20modelo%20completo>
[.https://www.youtube.com/watch?v=t2bfe4n8ETk](https://www.youtube.com/watch?v=t2bfe4n8ETk)
- Explicación del muestreo de Gibbs (ichi.pro)

