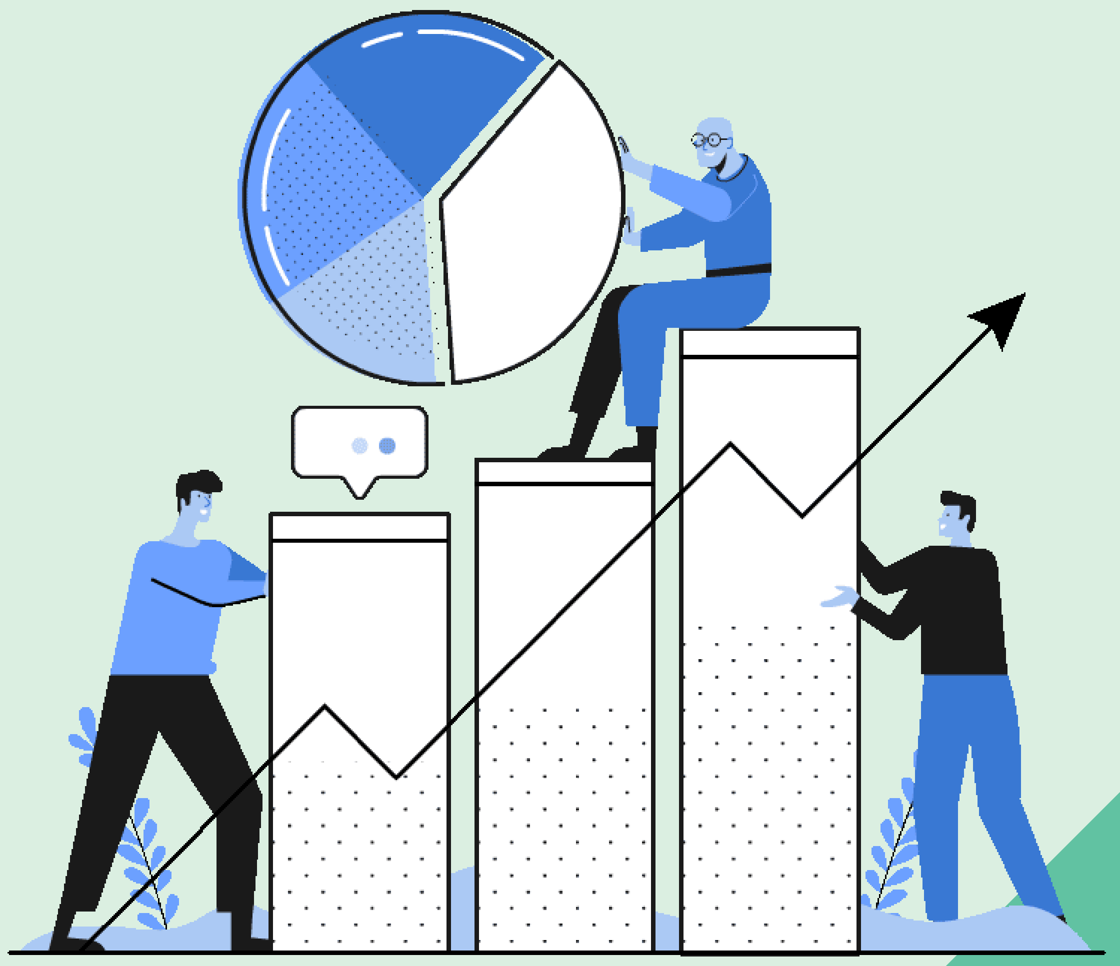
Regresión Bayesiana



- 1802780, Ramírez Medellín, Sahori Verónica
- 1937834, Flores Guerra, Irlanda Victoria
- 1806559, Treviño Elizondo, Melenie Anahí
- 1802294, Nuñes Márquez, Kevin Orlando

REGRESIÓN BAYESIANA

La aplicación del modelo bayesiano en modelos de regresión, sigue el esquema general de la estadística bayesiana:

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

TIPOS DE REGRESIÓN BAYESIANA

		Link	Inversa link
Modelo regresión	Función	$g(\mu) = \theta$	$\mu = g^{-1}(\theta)$
Normal	Identidad	μ	θ
Logístico	Logit	$\log\left(\frac{\mu}{1-\mu}\right)$	$\frac{\exp(\theta)}{1+\exp(\theta)}$
Probit	Normal inversa	$\Phi^{-1}(\mu)$	$\Phi(\theta)$
Poisson	logaritmo	$\log(\mu)$	$\exp(\theta)$
Gamma	Inversa	$-\frac{1}{\mu}$	$-\frac{1}{\theta}$
Binomial negativo	Logaritmo	$\log(1-\mu)$	$1-\exp(\theta)$

TEOREMA DE BAYES

Probabilidad Condicional $(A/B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(B) * P(A/B$

Teorema de Bayes $P(B_i/A) = \frac{P(B_i) * P(A/B_i)}{\sum_{i=1}^{n} P(B_i) * P(A/B_i)}$

Regla de probabilidad total $P(A) = \sum_{i=1}^{n} P(B_i) * P(A/B_i)$

Bi= Eventos no Observables

A=Evento Observable

P(Bi)=Priori

Priori-usar para actualizar información "previa"

Posteriori- la actualización de este previo usando la regla de Bayes que da la información después de haber visto los datos.

ACTUALIZACIÓN DE BAYES.

probabilidad basada en un evento que la afecta

PROCESO DE INFERENCIA

Bajo la inferencia bayesiana, las hipótesis científicas se expresan a partir de distribuciones de probabilidad formuladas a partir de la observación de los datos. Estas distribuciones dependen de magnitudes desconocidas denominadas parámetros $\boldsymbol{\theta}$, asignando a priori una distribución $p(\boldsymbol{\theta})$.

Para realizar la actualización de la distribución a priori $p(\boldsymbol{\theta})$, se utiliza la información que contienen los datos en relación a los parámetros expresada en la función de verosimilitud, que es proporcional a la distribución de los datos observados, dado los parámetros, y que se representa como $p(y/\boldsymbol{\theta})$.

El teorema de Bayes establece cómo actualizar la distribución a posteriori de $p(\theta)$. En este sentido, la distribución a posteriori es proporcional a la distribución a priori multiplicada por la función de verosimilitud, es decir:

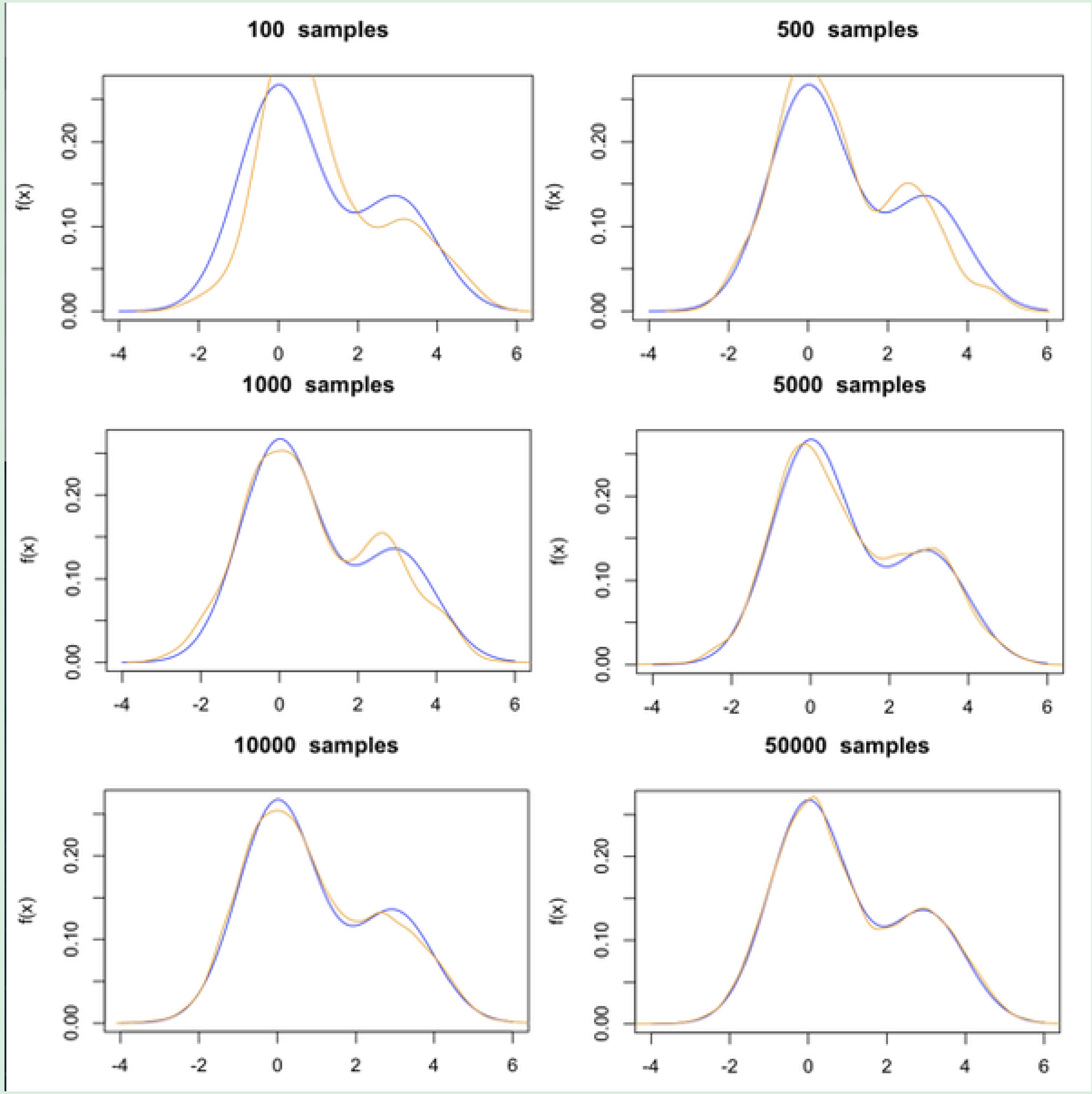
$$P(\theta/y) = \frac{P(\theta) * P(y/\theta)}{\int P(\theta) * P(y/\theta) d\theta}$$

o, alternativamente

posterior ∝ prior×verosimilitud

Los procedimientos de cálculo utilizados para la estimación de los modelos se engloban bajo **Markov Chain Monte Carlo**.

La integración **Monte Carlo** genera muestras de gran tamaño de la distribución de probabilidad de interés, a partir de las cuales evalúa los momentos de la distribución de probabilidad a **posteriori**.



Así, partiendo de la distribución a posteriori, podemos analizar cualquier característica (media, momentos, cuantiles, percentiles, zonas de mayor probabilidad,…), simplemente considerando su esperanza:

$$E(f(\theta/y)) = \frac{\int f(\theta)P(\theta) * P(y/\theta)d(\theta)}{\int P(\theta) * P(y/\theta)d\theta}$$

VENTAJAS

El uso de un marco de trabajo bayesiano tiene una serie de ventajas sobre la estadística frecuentista :

- Permite ajustar modelos complejos no abordables por métodos frecuentistas debido a las restricciones de estos modelos.

- Permite alcanzar estimaciones más exactas de los parámetros cuando el tamaño muestral es pequeño.

VENTAJAS

- La interpretación de los resultados es fácil y directa ya que indican la probabilidad de que un parámetro tome un cierto valor.
- Se pueden incluir medidas de incertidumbre, datos perdidos y diferentes niveles de variabilidad.
- Permite realizar propagaciones de error.
- Permite especificar las distribuciones de los parámetros (dependientes a su vez de otros parámetros) cuando a priori sabemos cómo se distribuyen (priors).
- Minimiza el uso de límites arbitrarios para tomar decisiones.

APLICACIONES

Las aplicaciones se encuentran en numerosas áreas de investigación como ingeniería, ciencias físicas, químicas, biológicas, etc.

Los modelos de regresión son usados para:

- -Descripción de datos
- -Estimación de parámetros
- -Predicción y estimación
- -Control

REGRESIÓN LINEAL BAYESIANA

REGRESIÓN LINEAL BAYESIANA

En un modelo de regresión simple

$$y_i = \underbrace{\alpha}_{\text{Intercepcción}} + \underbrace{\beta}_{\text{Pendiente de la recta}} x_i + \underbrace{\epsilon_i}_{\text{Error}}$$

se necesita definir la distribución conjunta a priori sobre **α**, **β** y **σ**

Dado que los residuos se distribuyen como una normal, $\boldsymbol{\epsilon i \sim N(0,\sigma)}$, la esperanza de la regresión será:

$$\mathbf{E(y)=\alpha+\beta x}$$

por lo tanto, se deduce que:

$$\mathbf{y \sim N(\alpha+\beta x,\sigma)}$$

VEROSIMILITUD

La función de verosimilitud será:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma}} exp[\frac{1}{\sigma^2}(y_i - (\alpha + \beta x_i))^2]$$

La conjunta, $\prod_{i=1}^{n} f(y_i)$ será por tanto una distribución normal multivariante de dimensión **n.**

Las distribuciones a posteriori de α y β son t-Student con n-2 grados de libertar y σ^2 se distribuye como una chi cuadrado inversa.

Distribuciones apriori independientes

Se especifican como distribuciones a priori independientes. Si suponemos:

$$\alpha \sim N(-\infty, \infty)$$

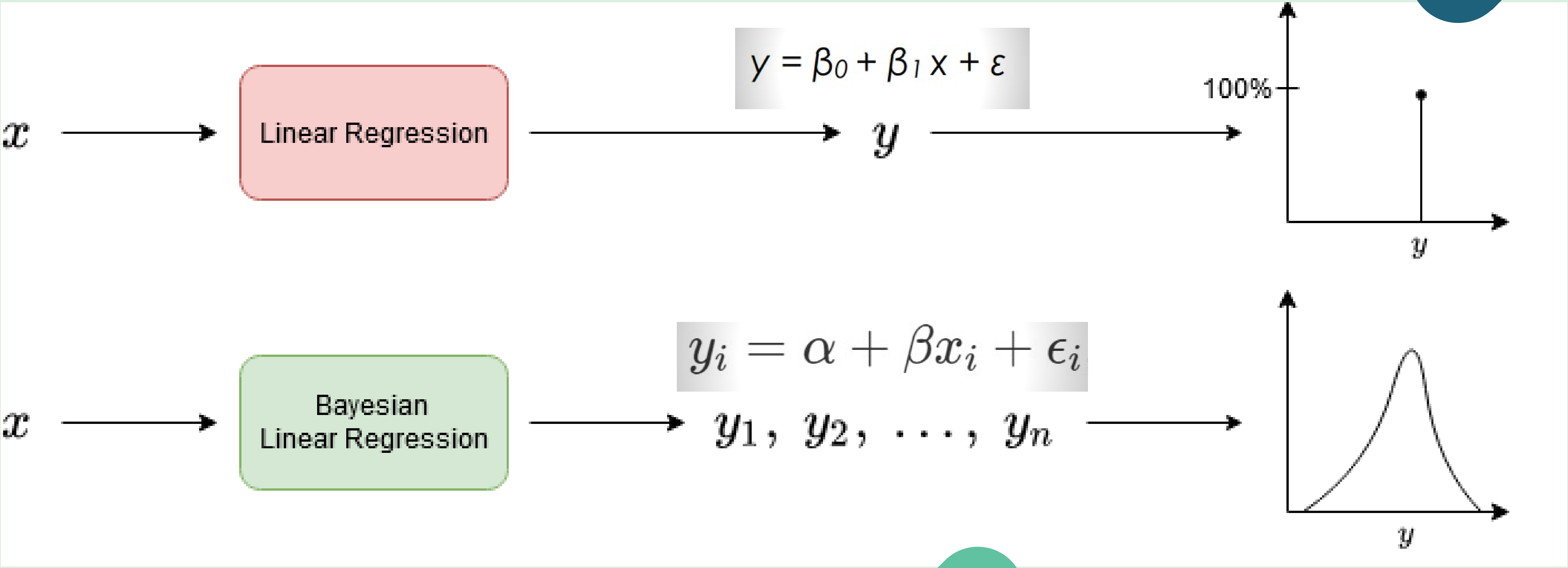
$$\beta \sim N(-\infty, \infty)$$

$$log(\sigma) \sim N(-\infty, \infty)$$

Por ser independientes, al multiplicarse producen una distribución conjunta a priori.

Se necesita considerar el logaritmo de la desviación estándar, dado que la varianza debe ser positiva.

COMPARACIÓN



EJEMPLO COMPARATIVO

REGRESION LINEAL
VS
REGRESION BAYESIANA

Se examinara una base de datos sobre la valoracion de viviendas en base a sus atributos pertenecientes a boston, EUA

OBJETIVO → **Saber cual modelo es mas eficiente**

```r
"Librerias que nos facilitaran la creacion de los modelos predictivos,
creacion de graficos y conjunto de datos econometricos"
library(ISLR);library(caret);library(arm);library(Ecdat);library(gridExtra)

data("Hedonic") #llamando a la base de datos
#Creando particiones de la base de datos para analisis posteriores
inTrain<-createDataPartition(y=Hedonic$tax,p=0.7, list=FALSE)
trainingset <- Hedonic[inTrain, ]
testingset <- Hedonic[-inTrain, ]
#mostrando la base de datos
str(Hedonic)
```

LIBRERIA ARM

Incluye modelos de regresión **bayesiana**, tiene la funcion "bayesglm" que implementa el modelo lineal generalizado bayesiano

DESCRIPCIÓN BASE DE DATOS

TASA DE CRIMINALIDAD

NUM. HABITACIONES

ANTIGUEDAD

TASA DE IMPUESTO

ACCESIBILIDAD

506 DATOS

15 COLUMNAS

```
#CREANDO MODELO DE REGRESION NORMAL LINEAL
ols.reg<-lm(tax~.,trainingset)
```

mv	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	blacks	lstat	townid
10.08580	0.00632	18.0	2.3099995	no	28.9444	43.2306	65.199951	1.40854	0.00000	296	15.29999	0.39690	-3.00074	1
9.98045	0.02731	0.0	7.0699997	no	21.9961	41.2292	78.899963	1.60283	0.69315	242	17.79999	0.39690	-2.39251	2
10.45450	0.02730	0.0	7.0699997	no	21.9961	51.6242	61.099976	1.60283	0.69315	242	17.79999	0.39283	-3.21165	2
10.41630	0.03237	0.0	2.1799984	no	20.9764	48.9720	45.799988	1.80207	1.09861	222	18.70000	0.39464	-3.52744	3

```r
"Hacemos nuestra predicción y compararemos los resultados con el conjunto de prueba utilizando:
correlación, estadísticas de resumen y el error absoluto medio"
ols.regTest<-predict.lm(ols.reg,testingset,interval = 'prediction',se.fit = T)
#correlacion
cor(testingset$tax,ols.regTest$fit[,1])
#estadisticos de resumen
summary(ols.regTest$fit[,1])
summary(trainingset$tax)
#error absoluto medio
MAE<-function(actual, predicted){
  mean(abs(actual-predicted))
}
MAE(ols.regTest$fit[,1], testingset$tax)
```

RESUMEN
MIN,MAX, CUARTILES

```
#estadisticos de resumen
summary(ols.regTest$fit[,1])
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
148.5   286.5   354.4   409.6   632.6   702.1
summary(trainingset$tax)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
187.0   278.5   330.0   405.7   666.0   711.0
```

MAE [1] 42.83572

COR [1] 0.9393213

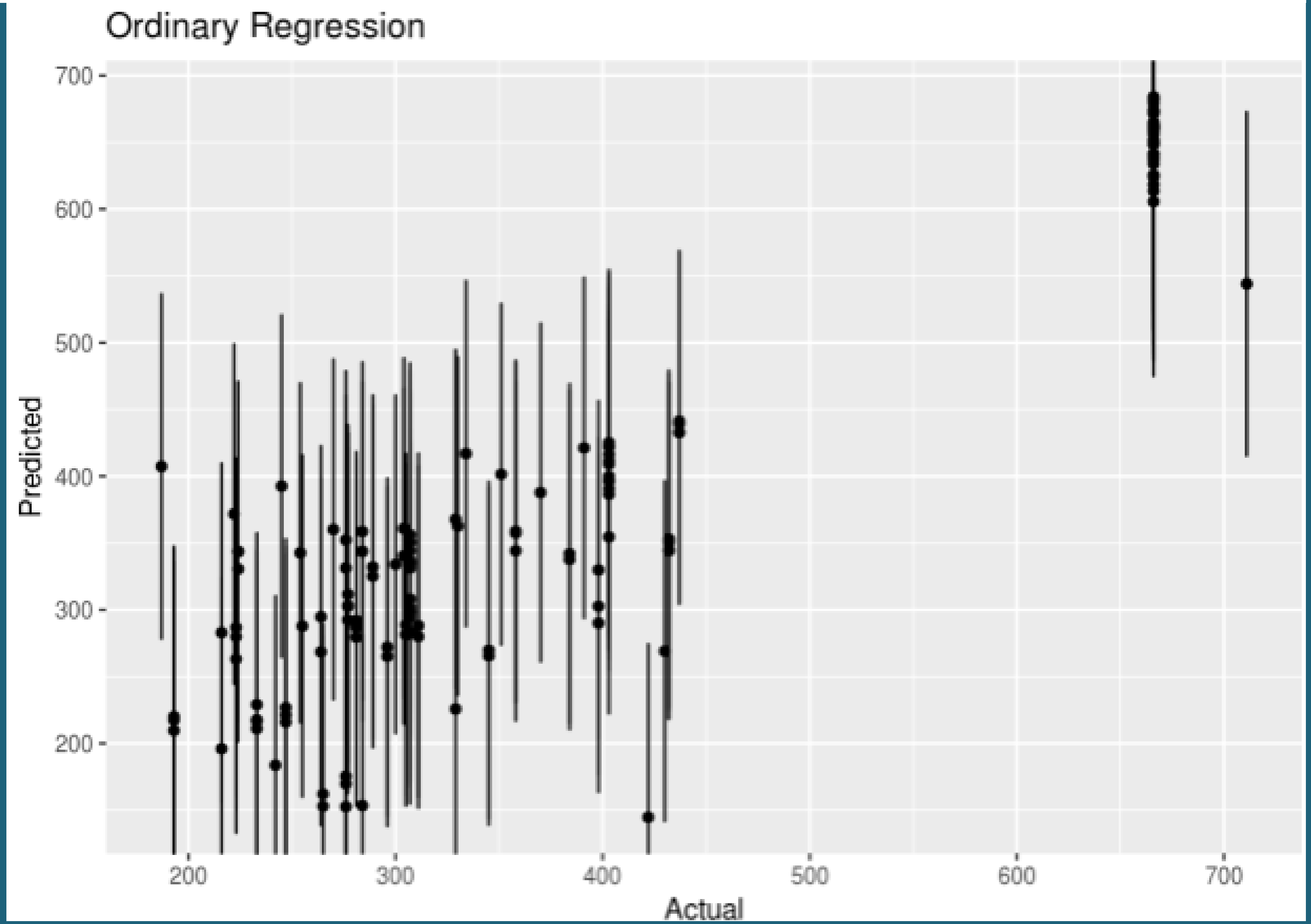
CONCLUSIÓN:

La correlación es buena.

Los resumen son similares y el error no es irrazonable.

```r
#variables utilizadas para formar el grafico
yout.ols <- as.data.frame(cbind(testingset$tax,ols.regTest$fit))
ols.upr <- yout.ols$upr
ols.lwr <- yout.ols$lwr
#creando grafico con las variables anteriores
p.ols <- ggplot(data = yout.ols, aes(x = testingset$tax, y = ols.regTest$fit[,1])) +
_geom_point() + ggtitle("Ordinary Regression") + labs(x = "Actual", y = "Predicted")
p.ols + geom_errorbar(ymin = ols.lwr, ymax = ols.upr)
```

GRÁFICA



```
#creando el modelo de regresion bayesiana
bayes.reg<-bayesglm(tax~.,family=gaussian(link=identity),trainingset,prior.df = Inf)
bayes.regTest<-predict.glm(bayes.reg,newdata = testingset,se.fit = T)
#correlacion
cor(testingset$tax,bayes.regTest$fit)
#resumen estadisticos
summary(bayes.regTest$fit)
#error cuadratico medio
MAE(bayes.regTest$fit, testingset$tax)
```

MAE [1] 42.83361

COR [1] 0.9393213

RESUMEN
MIN,MAX, CUARTILES

```
summary(bayes.regTest$fit)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
148.5   286.5   354.4   409.6   632.6   702.1
```

CONCLUSIÓN:

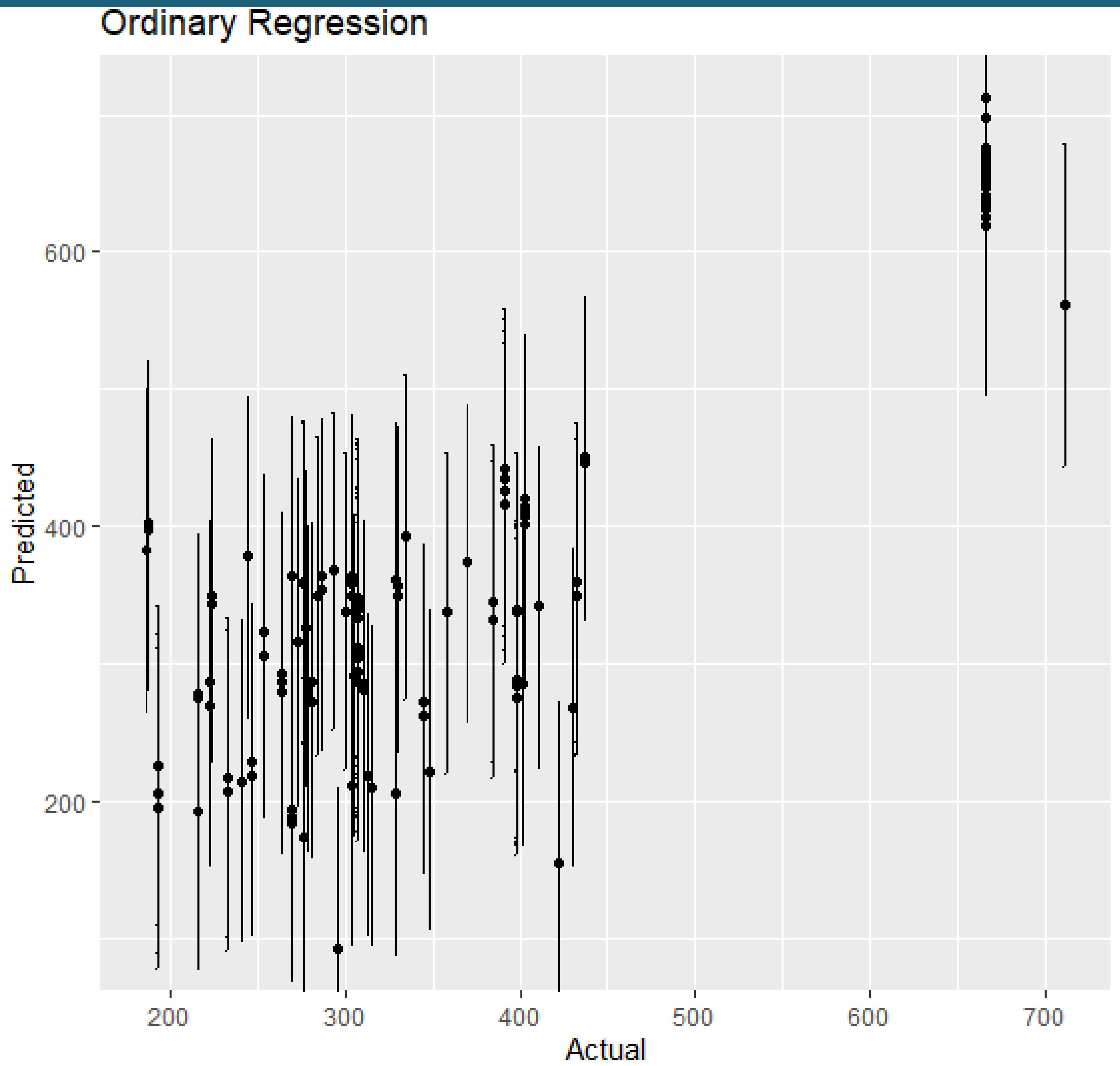
Los datos obtenidos por ambos modelos, son casi los mismos. Entonces, ¿Cuál es el beneficio de la regresión bayesiana?

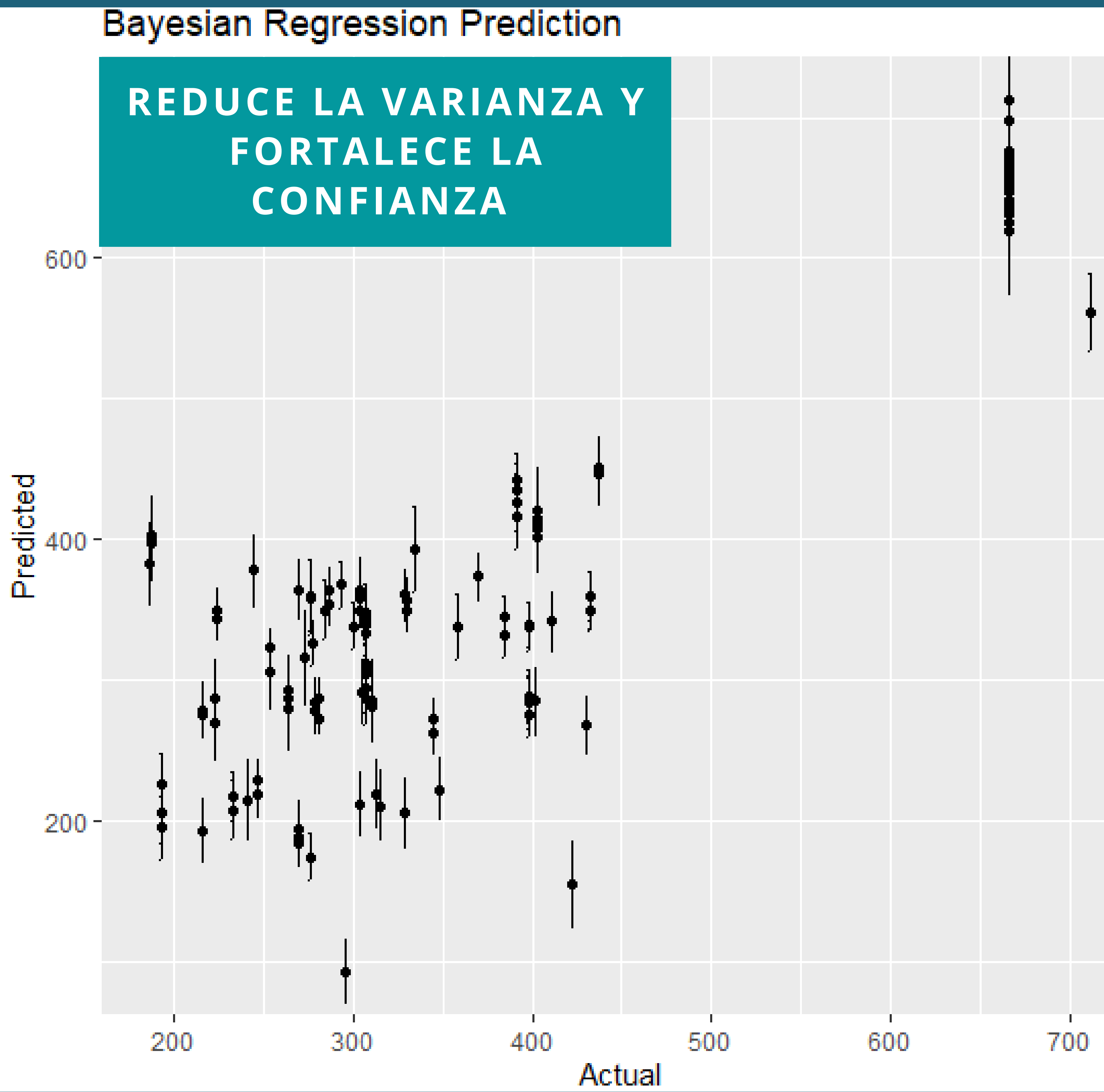
```r
#calculando intervalo de confianza para el bayesiano, pametros a utilizar en la grafica
yout.bayes <- as.data.frame(cbind(testingset$tax,bayes.regTest$fit))
names(yout.bayes) <- c("tax", "fit")
critval <- 1.96 #approx for 95% CI
bayes.upr <- bayes.regTest$fit + critval * bayes.regTest$se.fit
bayes.lwr <- bayes.regTest$fit - critval * bayes.regTest$se.fit
#creando grafico de regresion bayesiana
p.bayes <- ggplot(data = yout.bayes, aes(x = yout.bayes$tax, y = yout.bayes$fit)) + geom_point() +
_ggtitle("Bayesian Regression Prediction") + labs(x = "Actual", y = "Predicted")

#una vez creadas ambas graficas de las regresiones, se muestran en modo de comparacion
ols.plot <-  p.ols + geom_errorbar(ymin = ols.lwr, ymax = ols.upr)
bayes.plot <-  p.bayes + geom_errorbar(ymin = bayes.lwr, ymax = bayes.upr)
grid.arrange(ols.plot,bayes.plot,ncol=2)
```

La diferencia esta en el intervalo de confianza

COMPARATIVA





REDUCE LA VARIANZA Y FORTALECE LA CONFIANZA

Esto se debe a que en el enfoque bayesiano se calcula una distribución de parámetros a partir de una distribución posterior.

Para al final promediar los datos y obtener la predicción final que aparece en el gráfico.

BIBLIOGRAFIA

https://www.geeksforgeeks.org/types-of-regression-techniques/

https://www.geeksforgeeks.org/implementation-of-bayesian-regression/

https://economipedia.com/definiciones/analisis-de-regresion.html

https://bookdown.org/content/2274/modelos-con-variables-cualitativas.html#regresion-bayesiana

https://statswithr.github.io/book/bayesian-inference.html#continuous-variables-and-eliciting-probability-distributions

https://www.coursera.org/lecture/bayesian/bayes-rule-and-diagnostic-testing-5crO7

https://educationalresearchtechniques.com/2017/10/18/linear-regression-vs-bayesian-regression/