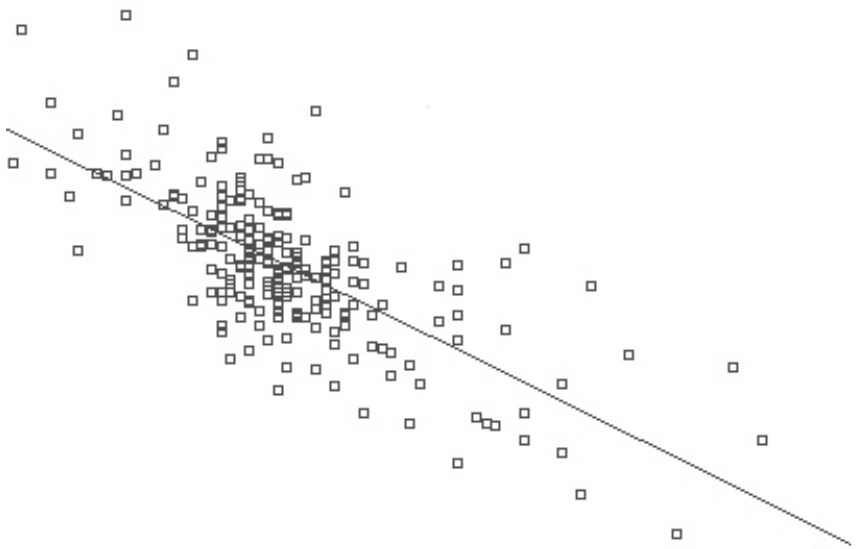
Regresión y modelos lineales



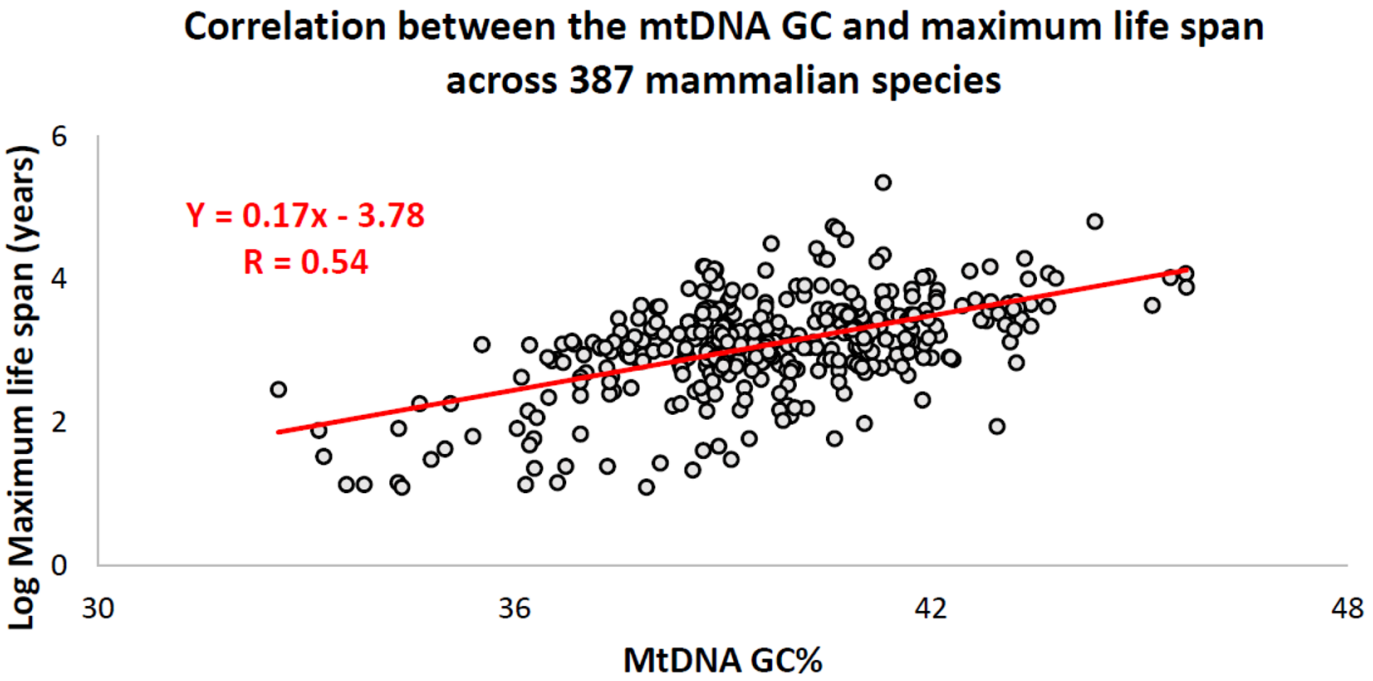
Mike Wiper

Departamento de Estadística

Universidad Carlos III de Madrid

Grado en Estadística y Empresa

Objetivo



Introducir el enfoque bayesiano a los modelos lineales.

Modelos lineales

Un modelo lineal tiene forma:

$$\mathrm{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

donde $\mathrm{y} = (y_1, ..., y_n)^T$ es un vector de "outputs", X es un $n \times k$ matriz de diseño, $\boldsymbol{\beta}$ es un vector de k parámetros y $\boldsymbol{\epsilon} \sim \text{Normal}(0, \sigma^2 \mathrm{I}_n)$ es un vector de errores aleatorios.

▸ MVN

Muchos modelos estandáres son modelos lineales:

Modelos lineales

Un modelo lineal tiene forma:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

donde $\mathbf{y} = (y_1, ..., y_n)^T$ es un vector de "outputs", $\mathbf{X}$ es un $n \times k$ matriz de diseño, $\boldsymbol{\beta}$ es un vector de k parámetros y $\boldsymbol{\epsilon} \sim \text{Normal}(0, \sigma^2 \mathbf{I}_n)$ es un vector de errores aleatorios.

▸ MVN

Muchos modelos estandáres son modelos lineales:

- Regresión lineal simple: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \boldsymbol{\beta} = (\beta_0, \beta_1)^T$$

Modelos lineales

Un modelo lineal tiene forma:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

donde $\mathbf{y} = (y_1, ..., y_n)^T$ es un vector de "outputs", $\mathbf{X}$ es un $n \times k$ matriz de diseño, $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^T$ es un vector de parámetros y $\boldsymbol{\epsilon} \sim \text{Normal}(0, \sigma^2 \mathbf{I}_n)$ es un vector de errores aleatorios.

Muchos modelos estandáres son modelos lineales:

- Regresión múltiple: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{k-1} x_{i,k-1} + \epsilon_i$.

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{i,k-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,k-1} \\ \vdots & \vdots & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,k-1} \end{pmatrix} \qquad \boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{k-1})^T$$

Modelos lineales

Un modelo lineal tiene forma:

$$\mathrm{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

donde $\mathrm{y} = (y_1, ..., y_n)^T$ es un vector de "outputs", X es un $n \times k$ matriz de diseño, $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^T$ es un vector de parámetros y $\boldsymbol{\epsilon} \sim \text{Normal}(0, \sigma^2 \mathrm{I}_n)$ es un vector de errores aleatorios.

Muchos modelos estandáres son modelos lineales:

- Análisis de varianza: $y_{ij} = \mu + \tau_j + \epsilon_{ij}$ for $i = 1, ..., n_j$, $j = 1, ..., g$.
- Análisis de covarianza: $\mathrm{y} = \mathrm{Z}\boldsymbol{\alpha} + \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ donde Z es una matriz de 0s y 1s como en análisis de varianza y X es un matriz de diseño como en regresión.
- Modelos lineales mezclados, modelos autoregresivos, ...

La función de verosimilitud para un modelo lineal

La verosimilitud es:

$$
\begin{aligned}
l(\boldsymbol{\mu}, \sigma^2|\mathrm{y}) &= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}(\mathrm{y}-\mathrm{X}\boldsymbol{\beta})^T(\mathrm{y}-\mathrm{X}\boldsymbol{\beta})\right) \\
&\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\left[\mathrm{y}^T\mathrm{y} - 2\boldsymbol{\beta}^T\mathrm{X}^T\mathrm{y} + \boldsymbol{\beta}^T\mathrm{X}^T\mathrm{X}\boldsymbol{\beta}\right]\right) \\
&\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\left[\mathrm{y}^T\mathrm{y} - 2\boldsymbol{\beta}^T(\mathrm{X}^T\mathrm{X})(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{y} + \boldsymbol{\beta}^T(\mathrm{X}^T\mathrm{X})\boldsymbol{\beta}\right]\right) \\
&\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\left[\mathrm{y}^T\mathrm{y} - 2\boldsymbol{\beta}^T(\mathrm{X}^T\mathrm{X})\hat{\boldsymbol{\beta}} + \boldsymbol{\beta}^T(\mathrm{X}^T\mathrm{X})\boldsymbol{\beta}\right]\right)
\end{aligned}
$$

donde $\hat{\boldsymbol{\beta}} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{y}$. Luego, completando la forma cuadratica, tenemos:

$$
\begin{aligned}
l(\boldsymbol{\mu}, \sigma^2|\mathrm{y}) &\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\left[\mathrm{y}^T\mathrm{y} - \hat{\boldsymbol{\beta}}^T(\mathrm{X}^T\mathrm{X})\hat{\boldsymbol{\beta}} + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\mathrm{X}^T\mathrm{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right]\right) \\
&\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\left[(\mathrm{y}-\mathrm{X}\hat{\boldsymbol{\beta}})^T(\mathrm{y}-\mathrm{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\mathrm{X}^T\mathrm{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right]\right)
\end{aligned}
$$

Inferencia frecuentista

El EMV de β es el estimador de mínimos cuadrados:

$$\hat{\beta} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{y}.$$

Típicamente se estima σ^2 con el estimador insesgado:

$$\begin{aligned}\hat{\sigma^2} &= \frac{\text{suma de errores al cuadrado}}{n - \text{número de parámetros}} = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \frac{1}{n-k}(\mathrm{y} - \mathrm{X}\hat{\beta})^T(\mathrm{y} - \mathrm{X}\hat{\beta})\end{aligned}$$

Ejemplo: regresión lineal simple

En este caso:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix}, \quad (\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{(n-1)s_x^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}.$$

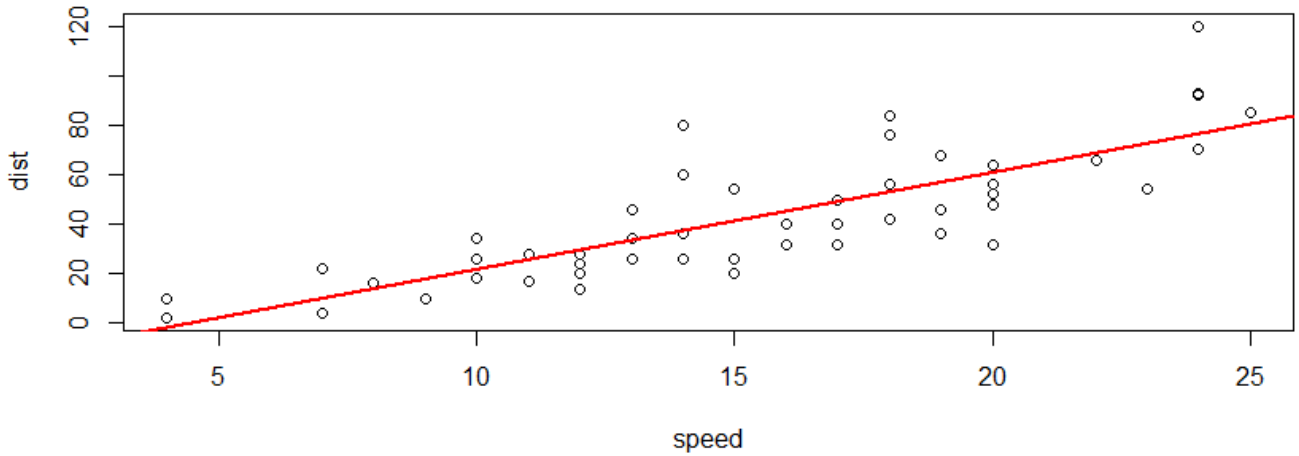
También, $\mathbf{X}^T\mathbf{y} = \begin{pmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$ y entonces, el EMV de $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ es

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Ejemplo: velocidades y tiempo necesario para parar de unos coches

```
help(cars)
data(cars)
plot(cars)
cor(cars$speed, cars$dist)
modelolineal <- lm(dist ~ speed, data=cars)
print(modelolineal)
abline(modelolineal,col='red',lwd=2)
anova(modelolineal)
xnuevo <- data.frame(20)
names(xnuevo) <- "speed"
predict(modelolineal,xnuevo,interval="confidence")
predict(modelolineal,xnuevo,interval="prediction")
```

Ejemplo: velocidades y tiempo necesario para parar de unos coches




```
Coefficients:
(Intercept)        speed
    -17.579        3.932

   fit       lwr       upr
61.06908 55.24729 66.89088
61.06908 29.60309 92.53507
```

Problemas

- Se necesitan más datos que parámetros.
- Dificultades con colinealidad.
- Se requiere que la matriz $X^T X$ sea invertible.

Para mitigar los problemas, típicamente se utiliza penalización:

- Regresión de arista o regresión de Tíjonov

$$\text{minimizar} : \sum_i \left(y_i - \sum_j \beta_j x_{ij} \right)^2 + \lambda \sum_j \beta_j^2$$

- Lasso

$$\text{minimizar} : \sum_i \left(y_i - \sum_j \beta_j x_{ij} \right)^2 + \lambda \sum_j |\beta_j|.$$

El enfoque bayesiano

Para el modelo bayesiano, es necesario definir distribuciones a priori para los parámetros, β, y la varianza del modelo, σ^2

Existen varias posibilidades:

- A priori conjugada
- A priori "no informativa"
- A priori semi-conjugada o no conjugada.

Una distribución a priori conjugada: la distribución normal-multivariante gamma

Sea $\phi = \frac{1}{\sigma^2}$ y definimos una distribución normal-multivariante gamma para $\boldsymbol{\beta}, \phi$:

$$\begin{aligned} \boldsymbol{\beta}|\phi &\sim \text{Normal}\left(\text{m}, \frac{1}{\phi}\text{V}\right) \\ \phi &\sim \text{Gamma}\left(\frac{a}{2}, \frac{b}{2}\right) \end{aligned}$$

La distribución conjunta es

$$f(\boldsymbol{\beta}, \phi) = \frac{\left(\frac{b}{2}\right)^{\frac{a}{2}}}{(2\pi)^{\frac{k}{2}}|V|^{\frac{1}{2}}\Gamma\left(\frac{a}{2}\right)}\phi^{\frac{a+k}{2}-1}\exp\left(-\frac{\phi}{2}\left[b + (\boldsymbol{\beta} - \text{m})^T\text{V}^{-1}(\boldsymbol{\beta} - \text{m})\right]\right)$$

La distribución marginal de β

Integrando con respeto a ϕ, tenemos:

$$\begin{aligned} f(\beta) &\propto \left(b + (\beta - \mathrm{m})^T \mathrm{V}^{-1} (\beta - \mathrm{m})\right)^{-\frac{a+k}{2}} \\ &\propto \left(1 + \frac{1}{a}(\beta - \mathrm{m})^T \left(\frac{b}{a}\mathrm{V}\right)^{-1} (\beta - \mathrm{m})\right)^{-\frac{a+k}{2}} \end{aligned}$$

que es una distribución t de Student multivariante escalada y no centrada.

La distribución t es bastante difícil de manejar pero observamos que es muy fácil simular de ello a través del Monte Carlo.

La distribución a posteriori

Dada la muestra, se puede demostar que la distribución a posteriori es normal-multivariante gamma con parámetros:

$$
\begin{aligned}
\mathrm{m}^* &= \left(\mathrm{V}^{-1} + \mathrm{X}^T\mathrm{X}\right)^{-1}\left(\mathrm{V}^{-1}\mathrm{m} + \mathrm{X}^T\mathrm{y}\right) \\
\mathrm{V}^* &= \left(\mathrm{V}^{-1} + \mathrm{X}^T\mathrm{X}\right)^{-1} \\
a^* &= a + n \\
b^* &= b + \mathrm{m}^T\mathrm{V}^{-1}\mathrm{m} + \mathrm{y}^T\mathrm{y} - \mathrm{m}^{*T}\mathrm{V}^{*-1}\mathrm{m}^*
\end{aligned}
$$

La distribución a posteriori

Dada la muestra, se puede demostar que la distribución a posteriori es normal-multivariante gamma con parámetros:

$$\begin{aligned} \mathrm{m}^* &= \left(\mathrm{V}^{-1} + \mathrm{X}^T\mathrm{X}\right)^{-1}\left(\mathrm{V}^{-1}\mathrm{m} + \mathrm{X}^T\mathrm{y}\right) \\ \mathrm{V}^* &= \left(\mathrm{V}^{-1} + \mathrm{X}^T\mathrm{X}\right)^{-1} \\ a^* &= a + n \\ b^* &= b + \mathrm{m}^T\mathrm{V}^{-1}\mathrm{m} + \mathrm{y}^T\mathrm{y} - \mathrm{m}^{*T}\mathrm{V}^{*-1}\mathrm{m}^* \end{aligned}$$

- Observamos que cuando el EMV existe, la media a posteriori es una media ponderada de la media a priori y el EMV:

$$E[\boldsymbol{\beta}|\mathrm{y}] = \left(\mathrm{V}^{-1} + \mathrm{X}^T\mathrm{X}\right)^{-1}\left(\mathrm{V}^{-1}\mathrm{m} + (\mathrm{X}^T\mathrm{X})\hat{\boldsymbol{\beta}}\right).$$

- Cuando el EMV no existe, todavía se puede calcular la media a posteriori. De hecho, coincide con el estimador de regresión arista en el caso $\mathrm{V} = \frac{1}{\lambda}\mathrm{I}$.

Predicción

Supongamos que queremos predicir una nueva observación $\tilde{Y} = \sum_{i=1}^{k} \beta_i \tilde{x}_i + \tilde{\epsilon}$.

Lo más sencillo es generar una muestra de la distribución a posteriori de $\boldsymbol{\beta}, \phi$ y después generar valores de ϵ de una distribución normal con precisiones iguales a los ϕ generados. Luego calculamos una muestra de $\tilde{Y}$ directamente.

Ejemplo

```
m <- c(0,1)
V <- 1000*diag(2)
Vinv <- solve(V)
a <- 1
b <- 1
X <- cars[,1]
n <- length(X)
X <- matrix(c(rep(1,n),X),nrow=n)
XTX <- t(X)%*%X
Vpost <- solve(XTX+Vinv)
y <- cars$dist
mpost <- Vpost%*%(Vinv%*%m+t(X)%*%y)
apost <- a+n
bpost <- b+t(m)%*%Vinv%*%m+t(y)%*%y-t(mpost)%*%solve(Vpost)%*%mpost
simuls <- 10000
library(MASS)
phi <- rgamma(simuls,apost/2,bpost/2)
beta <- matrix(rep(NA,simuls*2),nrow=simuls)
for (i in 1:simuls){
  beta[i,] <- mvrnorm(n = 1, mpost, Vpost/phi[i])
}
apply(beta,2,mean)
apply(beta,2,quantile,probs=c(0.025,0.975))
```

Ejemplo

Estimaciones de los parámetros y intervalos de credibilidad:

```
-17.434286    3.925996

              [,1]      [,2]
2.5%   -30.485629 3.113157
97.5%   -4.313447 4.733982
```

Predicciones de la media y de una nueva observación:

```
61.08564

    2.5%    97.5%
55.40086 66.80101
    2.5%    97.5%
30.48484 92.18414
```

A priori no informativa

Como anteriormente, supongamos que $f(\boldsymbol{\mu}, \phi) \propto \frac{1}{\phi}$. Luego la distribución a posteriori es una normal multivariante gamma con parámetros:

$$\begin{aligned} \mathrm{m}^* &= \left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\mathrm{y} = \hat{\boldsymbol{\beta}} \\ \mathrm{V}^* &= \left(\mathrm{X}^T\mathrm{X}\right)^{-1} \\ a^* &= a + n - k \\ b^* &= (\mathrm{y} - \mathrm{X}\hat{\boldsymbol{\beta}})^T(\mathrm{y} - \mathrm{X}\hat{\boldsymbol{\beta}}) \end{aligned}$$

En este caso, si la matriz $\mathrm{X}^T\mathrm{X}$ tiene inverso, la media a posteriori de $\boldsymbol{\beta}$ coincide con el EMV. Si no tiene inverso, (cuando $n \leq k$ o $\mathrm{X}^T\mathrm{X}$ no es invertible) la distribución a posteriori es impropia.

Ejemplo

Simulando de la a posteriori, tenemos estimaciones de los parámetros iguales (hasta el error de la simulación) a los EMV.

```
 -17.574671    3.930448

              [,1]      [,2]
2.5%   -30.780914 3.116529
97.5%   -4.501304 4.750646
```

Los intervalos predictivos son muy similares a las clásicas también.

```
61.03428

    2.5%    97.5%
55.39266 66.65916

    2.5%    97.5%
30.43020 91.23413
```

A priori semi conjugada

En problemas modernos, típicamente se suponen a prioris independientes para β, ϕ, por ejemplo $\beta \sim \text{Normal}(m, V)$. $\phi \sim \text{Gamma}\left(\frac{a}{2}, \frac{b}{2}\right)$.

En este caso, usamos el muestreo de Gibbs para implementar la inferencia.

```
m <- c(0,1)
V <- 1000*diag(2)
a <- 1
b <- 1
library(MCMCpack)
cc <- MCMCregress(dist ~ speed, data=cars, burnin = 1000,
                  mcmc = 10000, thin = 1,
                  beta.start = modelolineal$coefficients,
                  marginal.likelihood = c("none"))
summary(cc)
```

Ejemplo

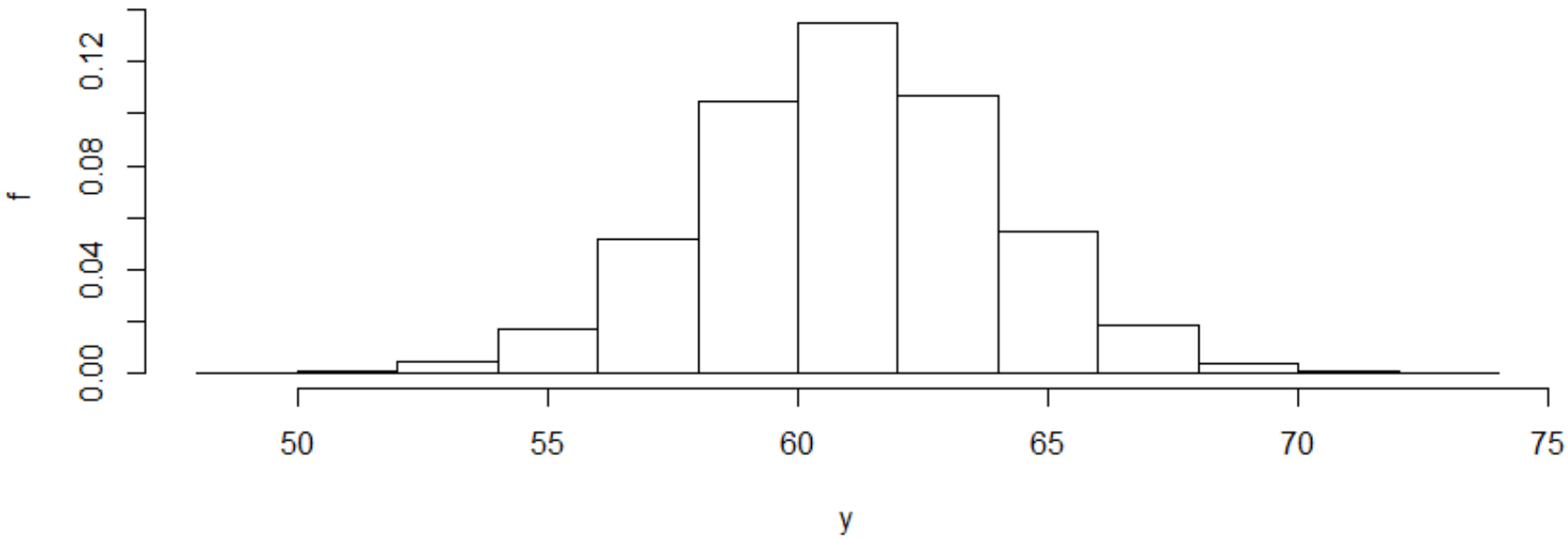
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
(Intercept)	-17.516	6.9119	0.069119	0.069119
speed	3.928	0.4265	0.004265	0.004265
sigma2	247.470	53.2858	0.532858	0.555348

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
(Intercept)	-31.089	-22.041	-17.421	-12.954	-4.087
speed	3.085	3.648	3.926	4.202	4.766
sigma2	164.940	209.887	240.105	277.494	371.629

Ejemplo: distribución predictiva de una nueva observación



Resumen y siguiente sesión

En esta sesión, hemos ilustrado el enfoque bayesiano a modelos lineales y regresión.

$$\begin{aligned} P(Y = y|\lambda) &= \frac{\lambda^y e^{-\lambda}}{y!} \\ \log \lambda &= \sum_{i=1}^{k} x_i \beta_i \end{aligned}$$

En la siguiente sesión, miraremos modelos lineales generalizados.

Apéndice: La distribución normal multivariante

Una variable multivariante $\mathrm{Y} = (Y_1, ..., Y_k)^T$ tiene una distribución normal multivariante con parámetros $\boldsymbol{\mu} \in \mathbb{R}^k$ y $\boldsymbol{\Sigma}$ una $(k \times k)$ matriz positiva semi definida si

$$f(\mathrm{y}) = \frac{1}{(2\pi)^{\frac{k}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathrm{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathrm{y} - \boldsymbol{\mu})\right) \quad \forall \mathrm{y} \in \mathbb{R}^k$$

La media es $E[\mathrm{Y}] = \boldsymbol{\mu}$ y la matriz de covarianzas es $V[\mathrm{Y}] = \boldsymbol{\Sigma}$.