Universidad
Rey Juan Carlos

Máster Universitario en Ingeniería de Sistemas de Decisión

Caso Práctico IV: Modelos Lineales

Autor:
Elisa Cascudo

Asignatura:
Modelización y Tratamiento de la Incertidumbre

13 Octubre 2025

Índice general

Índice de figuras

Capítulo 1

Selección de Variables

La base de datos que emplearemos en este practica, *Boston Housing Data*, contiene las siguientes variables:

- **CRIM:** representa el crimen per cápita por ciudad.
- **ZN:** proporción de zonas residenciales en un área determinada.
- **INDUS:** proporción de acres dedicada a negocios al por menor en la ciudad.
- **CHAS:** variable binaria (=1 si las vías cruzan el río y 0 en otro caso).
- **NOX:** concentración de óxido nítrico (partes por millón).
- **RM:** número medio de habitaciones por vivienda.
- **AGE:** proporción de edificios ocupados por sus propietarios, construidos antes de 1940.
- **DIS:** distancia ponderada a cinco centros de empleo en Boston.
- **RAD:** índice de accesibilidad a las autopistas radiales.
- **TAX:** valor total de la tasa de impuestos por 10.000 dólares.
- **PTRATIO:** ratio alumno-profesor por ciudad.
- **B:** valor definido como $1000(Bk-0{,}63)^2$, donde Bk es la proporción de afroamericanos en la ciudad.
- **LSTAT:** porcentaje de clase baja en la población.
- **MEDV:** valor medio de casas ocupadas por sus propietarios (en miles de dólares).

Para este caso practico, he decidido utilizar MEDV como variable a predecir. Por lo general el valor de las propiedades suele estar íntimamente conectado con otros factores de la zona, por ejemplo crimen o accesibilidad a autopistas, por lo que me ha parecido una variable interesante con gran potencial para relaciones lineales con otras variables.

El primer paso ha sido comprobar con que variables MEDV tiene una relación lineal, realizando gráficos de puntos que ilustrasen dichas relaciones. El código empleado en R para esta primera parte ha sido:

```
bhd <- read.table(file = "BHD.txt", header = FALSE)
colnames(bhd) <- c('CRIM','ZN','INDUS','CHAS','NOX','RM',
'AGE','DIS','RAD','TAX','PTRATIO','B','LSTAT','MEDV')
```

```
par(mfrow = c(3, 3))
for (x in colnames(bhd)) {
  if (x != "MEDV") {
    plot(bhd[[x]], bhd$MEDV,
         xlab = x,
         ylab = "MEDV",
         main = paste(x, "vs MEDV"))
  }
}
```

Como output, hemos obtenido las siguientes representaciones gráficas (al ser tan numerosas, he decidido colocarlas aquí en la practica en dos partes):

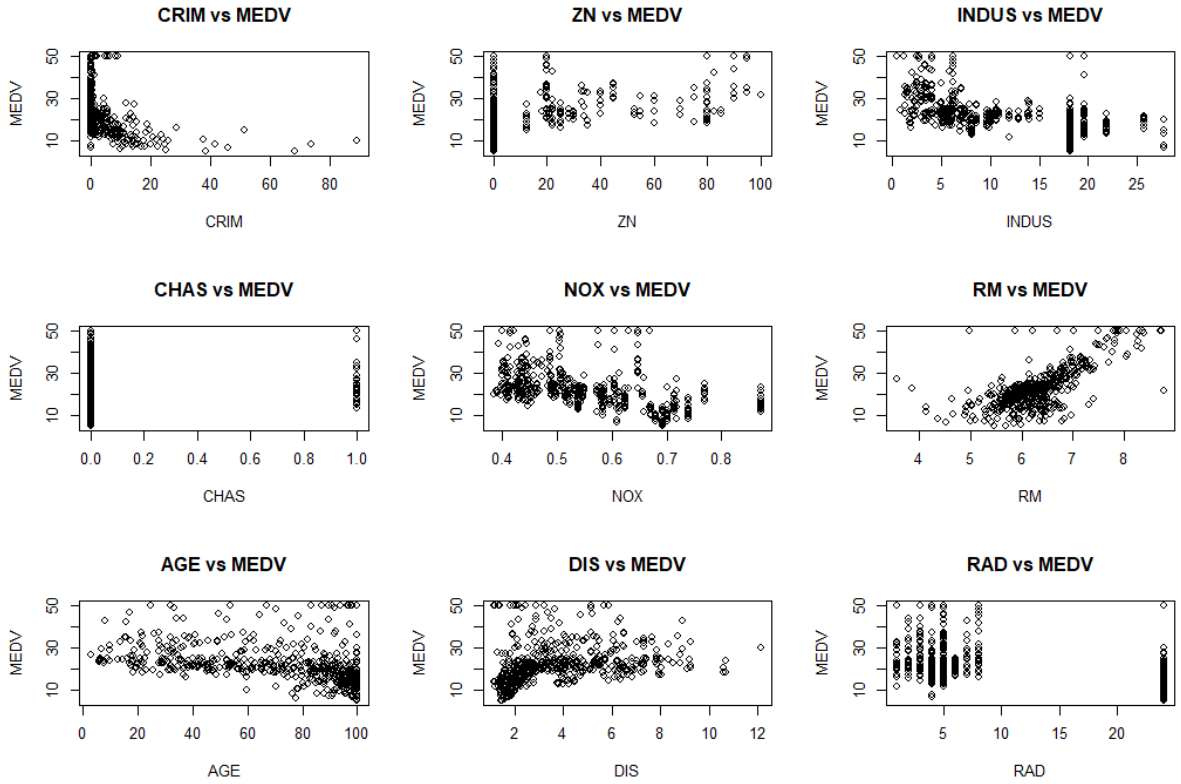

Figura 1.1: Relación entre MEDV y CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS y RAD ilustrada con un gráfico de puntos. Elaboración propia.

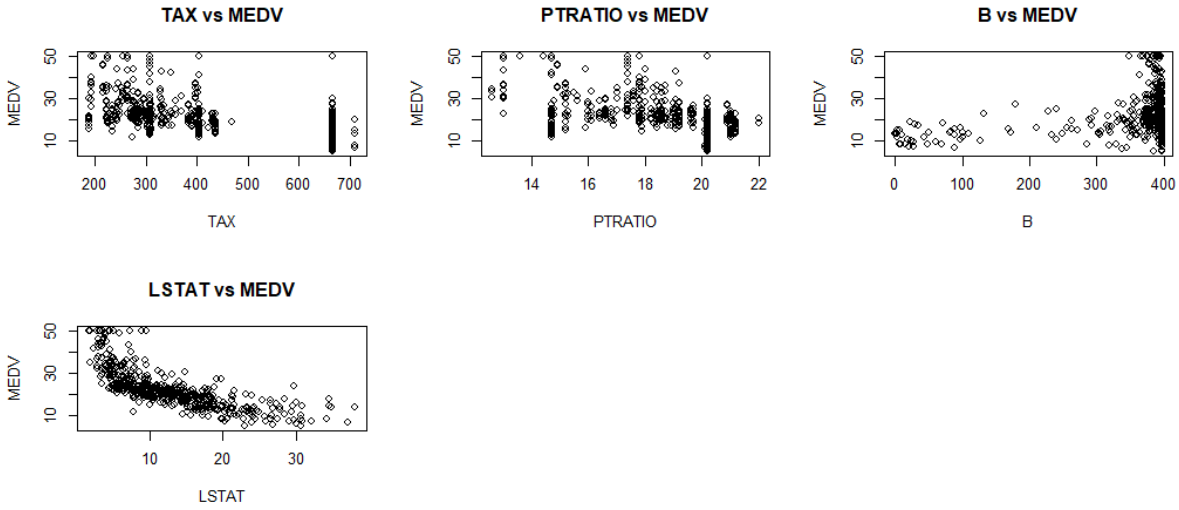

Figura 1.2: Relación entre MEDV y TAX, PTRATIO, B y LSTAT ilustrada con un gráfico de puntos. Elaboración propia.

1.1. Primera selección: a partir de los gráficos

Dando un primer vistazo a los graficos, aquellas variables que presentan una relación lineal directa con MEDV, y que se incluirán en un primer calculo del hiperplano de regresiónsin modificaciones son:

- **INDUS:** Presenta una relación lineal negativa.
- **CHAS:** Al ser una variable booleana, la dejaremos tal y como esta. Los valores de 0 y 1 presentan distinta relación con con los valores de MEDV, estando aquellas casas con vías que cruzan el río más acotado a un nivel medio y alto de valor de viviendas ocupadas por los propietarios.
- **NOX:** Presenta una relación lineal negativa.
- **RM:** Presenta una relación lineal positiva.
- **RAD:** Esta variable no presenta ni una tendencia lineal positiva ni negativa. Pero se hace interesante si la miramos desde un punto de vista de variables categóricas, lo cual nos podrá dar resultados interesantes mas adelante, al predecir posibles valores de MEDV.
- **TAX:** Aunque hay un vacío de valores entre aproximadamente 400.000 dólares y 600.000 dólares, y ruido entre los valores mas pequeños, he decidido incluirla, ya que la relación lineal negativa es bastante marcada.
- **PTRATIO:** Presenta una relación lineal negativa.

Antes de proceder con el análisis, es importante mencionar que se calculará un hiperplano de regresión, dado que el modelo que se construirá corresponde a una regresión multivariable. En este caso, la variable MEDV será predicha a partir de múltiples variables explicativas.

Aquellas variables a las que aplicaremos transformaciones son:

- **CRIM:** variable a la que aplicaremos la transformación Log(CRIM), ya que los valores están muy concentrados a la derecha.
- **DIS:** Aplicamos la misma transformación que a CRIM, Log(DIS)
- **LSTAT:** Y nos encontramos la misma situación con Log(LSTAT).

Ahora pasaremos a aquellas que desde un primer momento he elegido excluir desde un primer momento, y la justificación de esta decisión para cada una de ellas.

- **ZN:** Como se puede ver en la Figura 1.1, La variable ZN muestra una cierta relación lineal positiva con MEDV para valores superiores a cero. Sin embargo, se observa una fuerte concentración de observaciones en el valor cero, lo que indica un comportamiento categórico de la variable en esa zona. Debido a esta distribución desequilibrada y a la posible distorsión que podría generar en el modelo, decidí excluir ZN del análisis.
- **AGE:** La visualización de esta variable en la Figura 1.1 no aparenta ser ni negativa ni positiva, con ruido significativo a lo largo de todos los valores de AGE, por lo que he decidido excluirla.
- **B:** El problema que presenta esta variable se puede ver en la Figura 1.3. En una primera instancia los valores están concentrados a la izquierda, pero incluso después de aplicar a B una raíz cuadrada para tratar de estirar los datos, estos no presentan una transformación significativa. Por lo tanto, al no haber una relación lineal, se excluye del modelo.

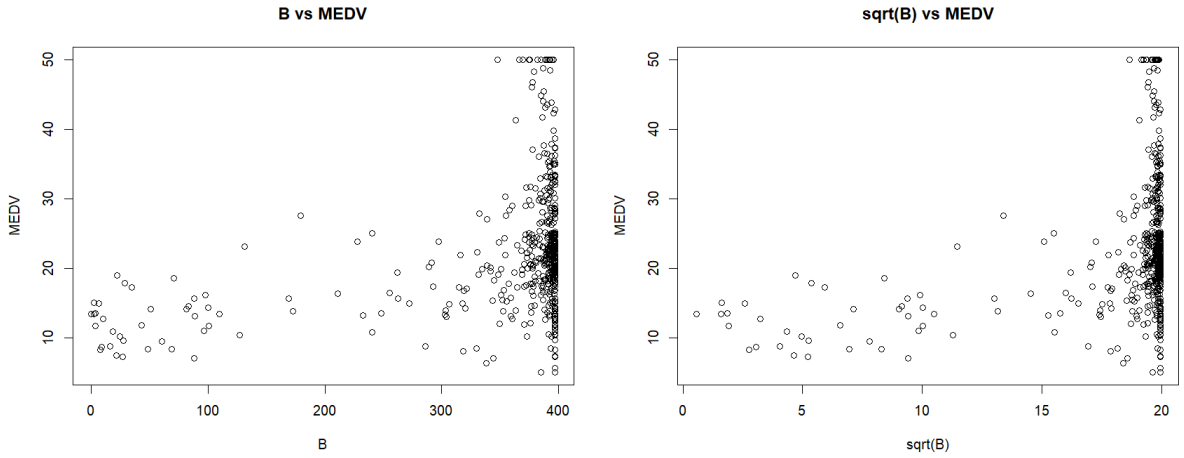

Figura 1.3: Representación del gráfico de puntos de B y raíz cuadrada de B frente a MEDV. Elaboración propia

1.2. Segunda selección: p-valor

El primer análisis realizado a partir de las gráficas proporciona información que nos permite dar el primer paso, pero ahora es necesario cribar las variables elegidas a través del p-valor obtenido al calcular el hiperplano de regresión. Si dicho valor se encuentra cerca de cero, las variables son significativas para el modelo que tenemos entre manos, y lo opuesto se deduce y si p-valor es muy próximo a 1.

El código empleado para esta comprobación ha sido:

```
fit <- lm(
  bhd$MEDV ~ log(bhd$CRIM) + bhd$INDUS + bhd$CHAS + bhd$NOX +
               bhd$RM + log(bhd$DIS) + bhd$RAD + bhd$TAX +
               bhd$PTRATIO + log(bhd$LSTAT),
  data = bhd,
  x = TRUE,
  y = TRUE
)

summary(fit)
```

Obteniendo como output:

```
Residuals:
     Min       1Q   Median       3Q      Max
-14.8194  -2.5442  -0.1788   2.1636  23.6983

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     63.223614   4.885038  12.942  < 2e-16 ***
log(bhd$CRIM)   -0.096159   0.238332  -0.403 0.686779
bhd$INDUS       -0.014575   0.055936  -0.261 0.794530
bhd$CHAS         2.642034   0.780079   3.387 0.000763 ***
bhd$NOX        -18.969841   3.642956  -5.207 2.82e-07 ***
bhd$RM           2.722164   0.379777   7.168 2.79e-12 ***
log(bhd$DIS)    -6.254223   0.705249  -8.868  < 2e-16 ***
bhd$RAD          0.213467   0.067960   3.141 0.001784 **
bhd$TAX         -0.012422   0.003314  -3.749 0.000199 ***
bhd$PTRATIO     -0.819896   0.112495  -7.288 1.25e-12 ***
log(bhd$LSTAT)  -9.184964   0.532423 -17.251  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.311 on 495 degrees of freedom
Multiple R-squared:  0.7847,    Adjusted R-squared:  0.7803
F-statistic: 180.4 on 10 and 495 DF,  p-value: < 2.2e-16
```

Figura 1.4: Output del calculo del primer hiperplano de regresión en R. Elaboración propia.

A partir de estos resultados, se pueden observar dos variables que en un principio parecían aportar información significativa para el modelo, pero que en realidad se deben excluir incluso tras su transformación logarítmica, debido a que presentan un p-valor muy próximo a 1.

- **CRIM:** El coeficiente de log(CRIM) no resulta significativo porque su efecto está fuertemente correlacionado con otras variables del modelo, como LSTAT, NOX, DIS e INDUS. Estas ya capturan gran parte de la información socioeconómica y ambiental que también refleja la tasa de criminalidad, por lo que CRIM no aporta variación explicativa adicional una vez que ellas están incluidas. En consecuencia, su influencia parcial sobre el valor medio de la vivienda se vuelve estadísticamente indistinguible de cero, aunque su relación individual con MEDV sea negativa.
- **INDUS:** En cuanto a apariencia esta variable parecía presentar una cierta relación lineal negativa, pero es verdad que contiene una gran cantidad de ruido al en los valores mas bajos que distorsionara el modelo. Además la relación lineal no es lo suficientemente significativa

Por lo tanto el modelo final se quedaria en:

```
fit <- lm(
  bhd$MEDV ~  bhd$CHAS + bhd$NOX + bhd$RM
            + log(bhd$DIS) + bhd$RAD + bhd$TAX +
              bhd$PTRATIO + log(bhd$LSTAT),
  data = bhd,
  x = TRUE,
  y = TRUE
)

summary(fit)
```

Con output:

```
Residuals:
     Min       1Q   Median       3Q      Max
-14.7468  -2.5049  -0.1682   2.1511  23.7126

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     63.709231   4.765990  13.367  < 2e-16 ***
bhd$CHAS         2.618404   0.774154   3.382 0.000776 ***
bhd$NOX        -19.625822   3.399139  -5.774 1.37e-08 ***
bhd$RM           2.733262   0.377688   7.237 1.75e-12 ***
log(bhd$DIS)    -6.128388   0.653878  -9.372  < 2e-16 ***
bhd$RAD          0.203615   0.054732   3.720 0.000222 ***
bhd$TAX         -0.012729   0.003014  -4.223 2.87e-05 ***
bhd$PTRATIO     -0.825315   0.110289  -7.483 3.32e-13 ***
log(bhd$LSTAT)  -9.230213   0.523710 -17.625  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.303 on 497 degrees of freedom
Multiple R-squared:  0.7846,    Adjusted R-squared:  0.7811
F-statistic: 226.2 on 8 and 497 DF,  p-value: < 2.2e-16
```

Figura 1.5: Output del calculo con las variables explicativas finales del hiperplano de regresión en R. Elaboración propia.

1.2.1. Análisis del hiperplano de regresión

El modelo final de regresión lineal múltiple presenta un ajuste satisfactorio, con un coeficiente de determinación R^2 ajustado de 0,7811, lo que indica que aproximadamente el 78 % de la variabilidad del valor medio de las viviendas ($MEDV$) es explicada por las variables incluidas en el modelo. El error estándar residual es de aproximadamente 4,3, reflejando una dispersión moderada en torno a la recta de regresión.

Todos los coeficientes resultan altamente significativos ($p < 0{,}001$), lo que sugiere que cada variable contribuye de manera relevante a la explicación de $MEDV$. La interpretación de los coeficientes estimados es coherente con la teoría económica y urbana:

- **CHAS** presenta un coeficiente positivo, indicando que las viviendas situadas junto al río tienen precios medios más altos.
- **NOX** tiene un efecto negativo: a mayor concentración de óxidos de nitrógeno, menor valor de las viviendas.
- **RM** muestra un efecto positivo, de modo que un mayor número de habitaciones se asocia con precios más elevados.
- **log(DIS)** tiene un coeficiente negativo, señalando que la mayor distancia a los centros de empleo reduce el valor medio.
- **RAD** presenta un efecto positivo leve, indicando que una mejor accesibilidad a autopistas tiende a incrementar el valor medio.
- **TAX** y **PTRATIO** muestran efectos negativos: los impuestos más altos y una peor ratio alumno/profesor se asocian con menores valores de vivienda.
- **log(LSTAT)** es el predictor con mayor efecto negativo: un incremento en la proporción de población de nivel socioeconómico bajo disminuye significativamente el valor medio de la vivienda.

En conjunto, el modelo obtenido refleja un buen comportamiento predictivo y una coherencia teórica elevada. El hiperplano estimado representa adecuadamente cómo las condiciones ambientales, educativas y socioeconómicas influyen en el valor medio de las viviendas en Boston, proporcionando una herramienta útil tanto para la interpretación como para la predicción.

Capítulo 2

Simulación de las distribuciones a posteriori de β_i y de la desviación típica del modelo

En esta sección queremos estimar la probabilidad de cada valor delos coeficientes del hiperplano, β_i, y de la varianza del modelo, σ^2. Posteriormente representaremos las distribuciones a posteriori obtenidas empleando histogramas. El objetivo al analizar estos histogramas es visualizar la forma de las distribución, para saber si presentan asimetrías, colas largas o algún sesgo que indique que el modelo no .Así podremos interpretar los intervalos de credibilidad visualmente.

En esta sección queremos estimar la probabilidad de cada valor de los coeficientes del hiperplano, β_i, y de la desviación estándar del modelo, σ^2. Posteriormente representaremos las distribuciones a posteriori obtenidas empleando histogramas. El objetivo al analizar estos histogramas es visualizar la forma de las distribuciones, para saber si presentan asimetrías, colas largas o algún sesgo que indique que el modelo no ha convergido adecuadamente o que existen problemas en la estimación.

Un comportamiento aproximadamente simétrico y con forma cercana a la normal en las distribuciones de los coeficientes β_i y de la incertidumbre σ sugiere una buena estabilidad en las estimaciones. En cambio, distribuciones muy asimétricas, multimodales (con solo una zona donde se concentra la mayor probabilidad) o con colas pronunciadas podrían revelar falta de convergencia, correlaciones fuertes entre parámetros o la necesidad de reconsiderar la especificación del modelo.

El código en R utilizado ha sido:

```
theta.sample=blinreg(fit$y,fit$x,5000)
#Perteneciendo blinreg al paquete LearnBayes

par(mfrow = c(3, 3))

hist(theta.sample$beta[,1], main = 'CHAS',          xlab = expression(beta[1]))
hist(theta.sample$beta[,2], main = 'NOX',           xlab = expression(beta[2]))
hist(theta.sample$beta[,3], main = 'RM',            xlab = expression(beta[3]))
hist(theta.sample$beta[,4], main = 'log(DIS)',      xlab = expression(beta[4]))
hist(theta.sample$beta[,5], main = 'RAD',           xlab = expression(beta[5]))
hist(theta.sample$beta[,6], main = 'TAX',           xlab = expression(beta[6]))
hist(theta.sample$beta[,7], main = 'PTRATIO',       xlab = expression(beta[7]))
hist(theta.sample$beta[,8], main = 'log(LSTAT)',    xlab = expression(beta[8]))
hist(theta.sample$sigma,    main = 'ERROR SD',      xlab = expression(sigma))
```

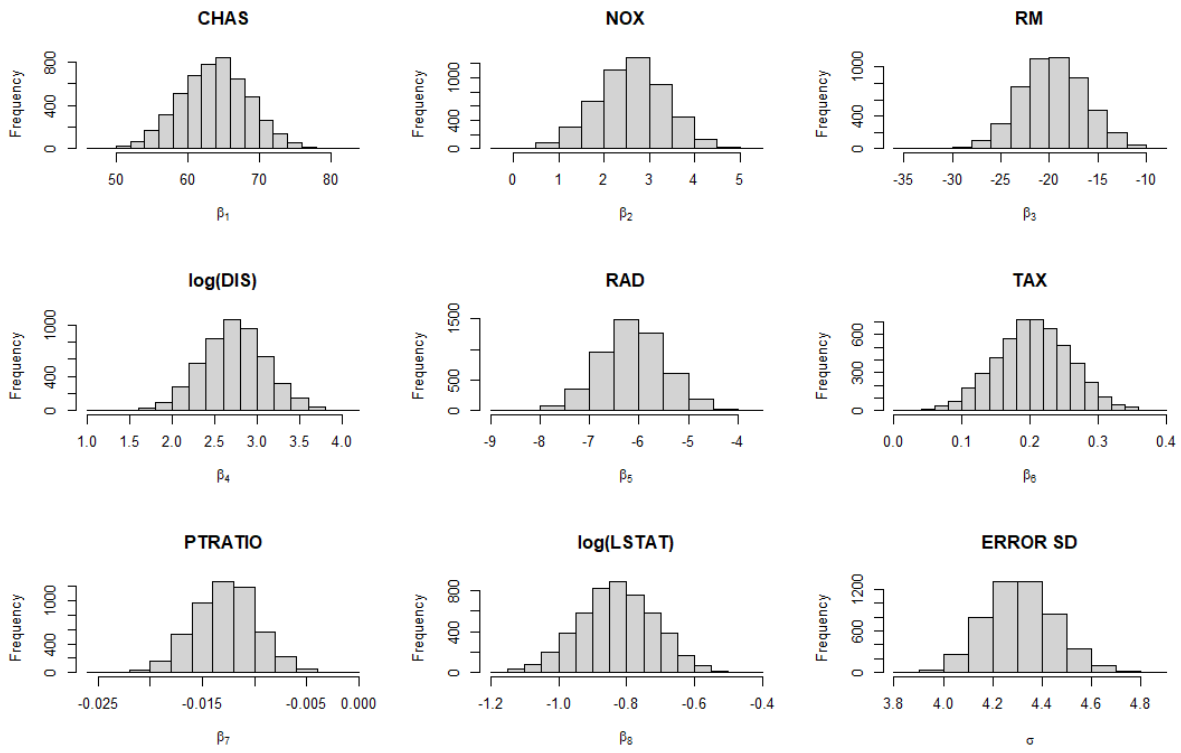

Figura 2.1: Histogramas de las distribuciones a posteriori delos coeficientes β_i y de la desviación típica, simuladas con blinreg. Elaboración propia.

En la Figura 2.1 se representan las distribuciones a posteriori de los coeficientes del modelo, β_i, y de la desviación estándar del error, σ. En general, todas las distribuciones presentan una forma aproximadamente normal y simétrica, centradas alrededor de un valor de media bien definido. Esto sugiere que las estimaciones son estables y no existen indicios de multimodalidad ni asimetrías marcadas. Asimismo, la dispersión relativamente pequeña indica una buena precisión en las estimaciones de los parámetros. No se aprecian colas alargadas ni sesgos marcados que pudieran sugerir la existencia de valores extremos o de incertidumbre elevada en torno a alguno de los parámetros.

- β_1 **(CHAS)**: La distribución es simétrica y claramente separada de cero, lo que confirma un efecto positivo y significativo de la proximidad al río sobre el valor medio de la vivienda.
- β_2 **(NOX)**: Presenta una distribución centrada en valores positivos, aunque algo más dispersa. La concentración de contaminantes tiene un efecto apreciable, aunque con cierta incertidumbre.
- β_3 **(RM)**: a distribución es estrecha y simétrica, indicando que un mayor número medio de habitaciones incrementa de forma consistente el valor de la vivienda.
- β_4 **(log(DIS))**: Distribución negativa, bien definida y con baja variabilidad. Muestra que la mayor distancia a los centros de empleo se asocia con precios medios menores.
- β_5 **(RAD)**: Distribución negativa y concentrada. Sugiere que una mayor accesibilidad a autopistas tiende a reducir ligeramente el valor medio.
- β_6 **(TAX)**: Distribución positiva, con efecto pequeño pero bien identificado sobre el valor medio de la vivienda.
- β_7 **(PTRATIO)**: Distribución centrada en valores negativos, indicando que una peor ratio alumno/profesor está asociada a precios de vivienda más bajos.

- β_8 **(log(LSTAT))**: Es la distribución más concentrada y claramente negativa, lo que muestra que el porcentaje de población de nivel socioeconómico bajo es el predictor más influyente en sentido negativo.
- σ **(Error SD)**: Presenta una distribución simétrica y unimodal alrededor de 4.4, lo que indica una estimación precisa y estable de la incertidumbre residual del modelo.

En conjunto, las distribuciones a posteriori obtenidas muestran que el modelo es estable y presenta una convergencia adecuada. Los signos de los coeficientes son coherentes con la teoría y los valores obtenidos en la Figura 1.5 lo que confirma la consistencia y fiabilidad de las estimaciones bayesianas de los parámetros.

Capítulo 3

Resumen de los percentiles

A continuación, voy a realizar el resumen de los percentiles correspondientes a la distribución a posteriori del modelo. Este análisis permite obtener una estimación de los intervalos de credibilidad de los parámetros, proporcionando una medida directa de la incertidumbre asociada a las estimaciones bayesianas. Además, los valores de estos percentiles permiten contrastar y confirmar los resultados obtenidos previamente en la estimación del hiperplano de regresión representado en la Figura 1.5.

Para que las conclusiones extraídas hasta este punto puedan considerarse consistentes, es necesario que las estimaciones puntuales derivadas de la distribución a posteriori (por ejemplo, la mediana a posteriori) sean coherentes con los coeficientes obtenidos en el modelo del hiperplano de regresión, y que los intervalos de credibilidad reflejen una variabilidad razonable. Ciertas diferencias entre ambos enfoques son esperables, dado que el procedimiento bayesiano se basa en simulaciones.

Para ello, hemos empleado el codigo:

```
apply(theta.sample$beta,2,quantile,c(0.05,0.5,0.95))
quantile(theta.sample$sigma, c(.05, .5, .95))
```

Y el output obtenido ha sido:

```
> apply(theta.sample$beta,2,quantile,c(0.05,0.5,0.95))
    X(Intercept) Xbhd$CHAS  Xbhd$NOX  Xbhd$RM Xlog(bhd$DIS)
5%      55.81620  1.305466 -24.96552 2.104358     -7.175359
50%     63.87304  2.616146 -19.63612 2.732742     -6.124820
95%     71.72597  3.856438 -14.00178 3.362375     -5.031436
     Xbhd$RAD     Xbhd$TAX Xbhd$PTRATIO Xlog(bhd$LSTAT)
5%  0.1144076 -0.017708451   -1.0141002      -10.089682
50% 0.2047905 -0.012723355   -0.8275163       -9.221365
95% 0.2942984 -0.007841869   -0.6461080       -8.381134
> quantile(theta.sample$sigma, c(.05, .5, .95))
      5%      50%      95%
4.086950 4.307090 4.549812
```

Figura 3.1: Percentiles 5 %, 50 % y 95 % obtenidos de la distribución a posteriori de los coeficientes β_i y de la desviación estándar del modelo

Si comparamos las medianas a posteriori obtenidas para cada coeficiente β_i en la Figura 3.1 con los valores del hiperplano de regresión de la Figura 1.5, nos encontramos con valores muy similares, corroborando la coherencia del modelo.

Capítulo 4

Estimación de media y predicción de MEDV

4.1. Estimación de la media de MEDV

Nuestro objetivo en esta sección es estimar la media del valor de la variable de MEDV a partir de la distribución a posteriori simulada del resto de las variables explicativas del modelo, y posteriormente realizar una predicción de su valor según ciertos parámetros para cada una delas demás variables.

Para ello estudiare como elementos fuera de las características de la casa (numero de cuartos, impuestos a pagar, nivel económico de sus habitantes, etc) afectan el valor medio de las casas ocupadas. Principalmente las variables que conciernen a la localización geográfica y accesibilidad de las viviendas. Aquellos a los que aplicaremos variabilidad serán:

- **CHAS:** que representa si las vías cruzan el rio o no a través de variables booleanas 0 y 1.
- **log(DIS):** al corresponder a la distancia ponderada a cinco centros de empleo de Boston distintos.
- **RAD:** como indicador de nivel de accesibilidad a las autopistas radiales.

Para los demás ítems, mantendremos el valor de la media obtenido en la Figura 1.5. Por lo tanto los conjuntos covariables a definir tendrán la siguiente forma en R:

```
#Definimos las variables que tomaran sus valores medios:
NOX      <-  2.6184
RM       <-  1.7332
TAX      <- -0.0127
PTRATIO  <- -0.82531
logLSTAT <- -9.2302

#Definimos aquellos valores que tendran variabilidad y serviran para la comparacion:
CHAS0 <- 0; CHAS1 <- 1
logDIS_low  <- 2.25
logDIS_high <- 3.25
RAD_low  <- -7
RAD_high <- -5

cov1 <- c(1, CHAS0, NOX, RM, logDIS_low,  RAD_low,  TAX, PTRATIO, logLSTAT)
cov2 <- c(1, CHAS1, NOX, RM, logDIS_low,  RAD_low,  TAX, PTRATIO, logLSTAT)
cov3 <- c(1, CHAS0, NOX, RM, logDIS_high, RAD_low,  TAX, PTRATIO, logLSTAT)
```

```
cov4 <- c(1, CHAS1, NOX, RM, logDIS_high, RAD_low,  TAX, PTRATIO, logLSTAT)
cov5 <- c(1, CHAS0, NOX, RM, logDIS_low,  RAD_high, TAX, PTRATIO, logLSTAT)
cov6 <- c(1, CHAS1, NOX, RM, logDIS_low,  RAD_high, TAX, PTRATIO, logLSTAT)
cov7 <- c(1, CHAS0, NOX, RM, logDIS_high, RAD_high, TAX, PTRATIO, logLSTAT)
cov8 <- c(1, CHAS1, NOX, RM, logDIS_high, RAD_high, TAX, PTRATIO, logLSTAT)

X1 <- rbind(cov1,cov2,cov3,cov4,cov5,cov6,cov7,cov8)

## Estimaci n de la media esperada E[MEDV | X] desde la posterior
mean.draws=blinregexpected(X1,theta.sample)

## Histos de las distintas distribuciones
par(mfrow = c(2,4), mar=c(4,4,3,1))
hist(mean.draws[,1], main="CHAS=0, logDIS=2.25, RAD=-7",  xlab="E[MEDV]")
hist(mean.draws[,2], main="CHAS=1, logDIS=2.25, RAD=-7",  xlab="E[MEDV]")
hist(mean.draws[,3], main="CHAS=0, logDIS=3.25, RAD=-7",  xlab="E[MEDV]")
hist(mean.draws[,4], main="CHAS=1, logDIS=3.25, RAD=-7",  xlab="E[MEDV]")
hist(mean.draws[,5], main="CHAS=0, logDIS=2.25, RAD=-5",  xlab="E[MEDV]")
hist(mean.draws[,6], main="CHAS=1, logDIS=2.25, RAD=-5",  xlab="E[MEDV]")
hist(mean.draws[,7], main="CHAS=0, logDIS=3.25, RAD=-5",  xlab="E[MEDV]")
hist(mean.draws[,8], main="CHAS=1, logDIS=3.25, RAD=-5",  xlab="E[MEDV]")
```

El gráfico de histogramas final obtenido ha sido:

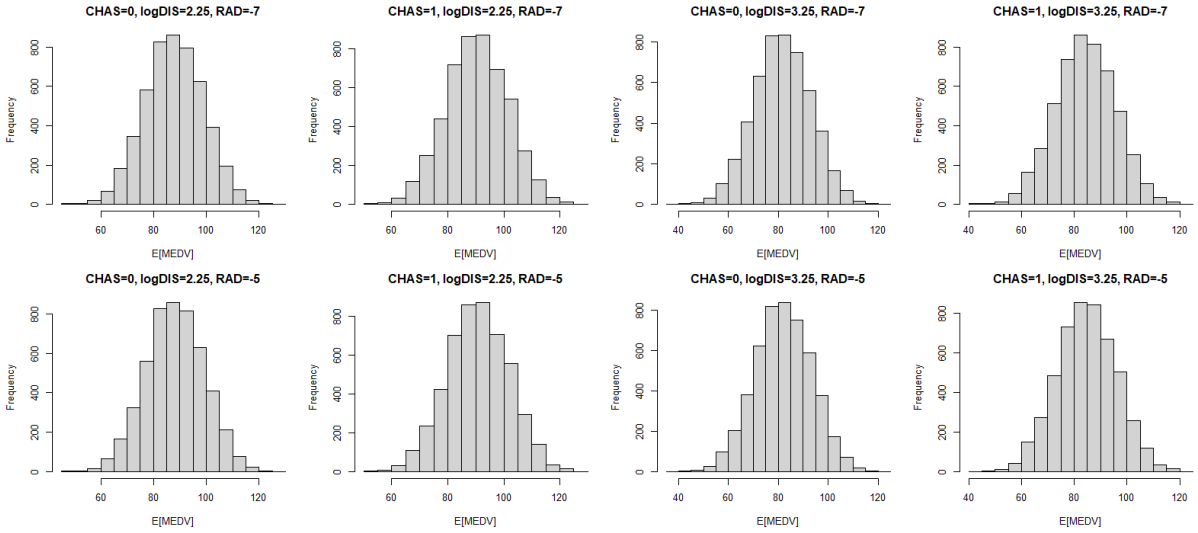

Figura 4.1: Calculo de la estimación de MEDV según distintos valores de CHAS, DIS y RAD. Elaboración propia.

En la Figura 4.1 se muestran las distribuciones a posteriori de la media esperada de `MEDV` para los ocho escenarios definidos. En todos los casos, las distribuciones presentan una forma aproximadamente normal y unimodal, con una dispersión similar entre escenarios. Esto indica que las simulaciones obtenidas mediante el modelo bayesiano son estables y que la incertidumbre posterior asociada a la media esperada de `MEDV` es comparable en todos los supuestos analizados. Asimismo, la ausencia de colas alargadas o distribuciones multimodales sugiere una buena convergencia de las cadenas y una especificación adecuada del modelo.

Al comparar los escenarios, se observa que al pasar de `CHAS = 0` a `CHAS = 1`, manteniendo el resto de variables constantes, las distribuciones se desplazan hacia valores mayores de `E[MEDV]`, lo que indica un

efecto positivo de la variable `CHAS`. Este resultado confirma que las viviendas situadas junto al río tienden a presentar un mayor valor medio de mercado.

Por otro lado, al aumentar el valor de `log(DIS)` de 2.25 a 3.25, las distribuciones se desplazan visiblemente hacia la izquierda en todos los escenarios. Este comportamiento muestra que la distancia a los centros de empleo tiene un efecto negativo más pronunciado sobre el valor medio de las viviendas. Es decir, a medida que las viviendas se encuentran más alejadas de los principales centros de empleo, su valor medio disminuye.

En cuanto a la variable `RAD`, al pasar de -7 a -5 se aprecia también un desplazamiento hacia la izquierda, aunque de menor magnitud que en el caso de `log(DIS)`. Este resultado indica que una mayor accesibilidad a las autopistas radiales se asocia con una ligera reducción en el valor medio de las viviendas, aunque su efecto es más moderado que el de las otras covariables analizadas.

En términos generales, el análisis muestra que el factor con mayor influencia sobre `MEDV` es la distancia a los centros de empleo, seguido por la cercanía al río y, en menor medida, la accesibilidad a las autopistas. En conjunto, los resultados son coherentes con la lógica urbana. Estar cercano a los centros de empleo, y consecuentemente a los centros de actividad económica, es el factor de mayor valor para las comunidades de Boston, lo cual tiene sentido ya que es fundamental para la supervivencia económica y material de los habitantes de cualquier ciudad. Permite obtener medios económicos más fácilmente, estar cercano a bienes de necesidad básica, e incluso mayor acceso a espacios de socialización, que se suelen concentrar en este tipo de núcleos de empleo.

4.2. Predicción de MEDV

Para predecir MEDV, utilizaremos las mismas covariables definidas en el apartado anterior, pero esta vez para predecir una estimación de MEDV, no calcularla. El código empleado para este caso ha sido:

$$
\begin{aligned}
\text{NOX} &\leftarrow 2.6184 \\
\text{RM} &\leftarrow 1.7332 \\
\text{TAX} &\leftarrow -0.0127 \\
\text{PTRATIO} &\leftarrow -0.82531 \\
\text{logLSTAT} &\leftarrow -9.2302
\end{aligned}
$$

$$
\begin{aligned}
&\text{CHAS0} \leftarrow 0;\ \text{CHAS1} \leftarrow 1 \\
&\text{logDIS\_low} \leftarrow 2.25 \\
&\text{logDIS\_high} \leftarrow 3.25 \\
&\text{RAD\_low} \leftarrow -7
\end{aligned}
$$

```
cov1 <- c(1, CHAS0, NOX, RM, logDIS_low , RAD_low, TAX, PTRATIO, logLSTAT)
cov2 <- c(1, CHAS1, NOX, RM, logDIS_low , RAD_low, TAX, PTRATIO, logLSTAT)
cov3 <- c(1, CHAS0, NOX, RM, logDIS_high , RAD_low, TAX, PTRATIO, logLSTAT)
cov4 <- c(1, CHAS1, NOX, RM, logDIS_high , RAD_low, TAX, PTRATIO, logLSTAT)
cov5 <- c(1, CHAS0, NOX, RM, logDIS_low , RAD_high, TAX, PTRATIO, logLSTAT)
cov6 <- c(1, CHAS1, NOX, RM, logDIS_low , RAD_high, TAX, PTRATIO, logLSTAT)
cov7 <- c(1, CHAS0, NOX, RM, logDIS_high , RAD_high, TAX, PTRATIO, logLSTAT)
cov8 <- c(1, CHAS1, NOX, RM, logDIS_high , RAD_high, TAX, PTRATIO, logLSTAT)

X1 <- rbind(cov1 ,cov2 ,cov3 ,cov4 ,cov5 ,cov6 ,cov7 ,cov8)
```

```
mean.draws <- blinregpred(X1, theta.sample)

par(mfrow = c(2,4), mar=c(4,4,3,1))
hist(mean.draws[,1], main="CHAS=0, logDIS=2.25, RAD=-7", xlab="E[MEDV]")
hist(mean.draws[,2], main="CHAS=1, logDIS=2.25, RAD=-7", xlab="E[MEDV]")
hist(mean.draws[,3], main="CHAS=0, logDIS=3.25, RAD=-7", xlab="E[MEDV]")
hist(mean.draws[,4], main="CHAS=1, logDIS=3.25, RAD=-7", xlab="E[MEDV]")
hist(mean.draws[,5], main="CHAS=0, logDIS=2.25, RAD=-5", xlab="E[MEDV]")
hist(mean.draws[,6], main="CHAS=1, logDIS=2.25, RAD=-5", xlab="E[MEDV]")
hist(mean.draws[,7], main="CHAS=0, logDIS=3.25, RAD=-5", xlab="E[MEDV]")
hist(mean.draws[,8], main="CHAS=1, logDIS=3.25, RAD=-5", xlab="E[MEDV]")
```

Y los histogramas obtenidos:

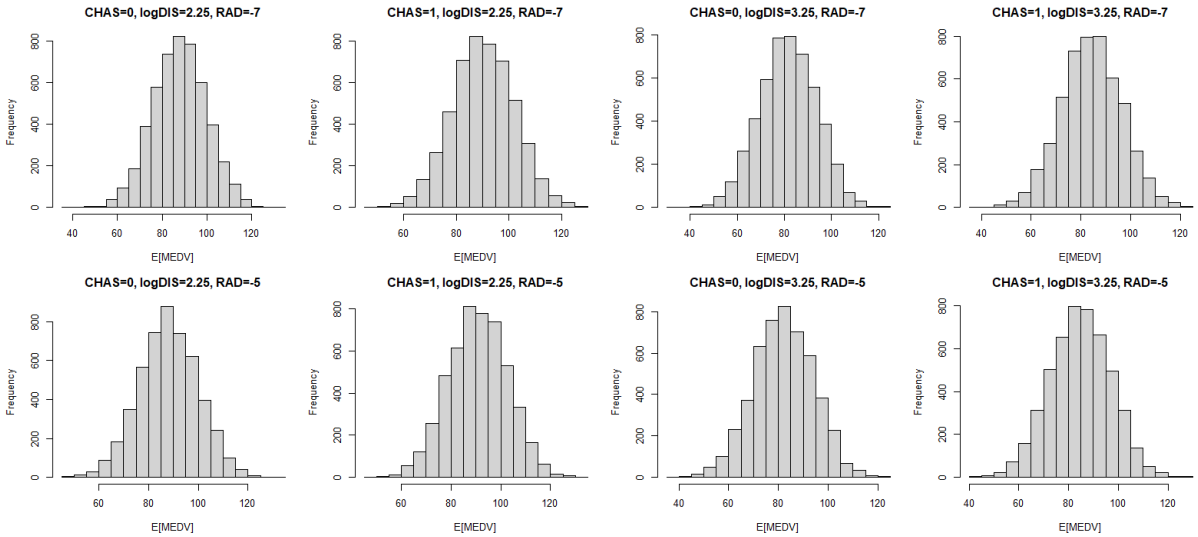

Figura 4.2: Predicción de la estimación de MEDV según distintos valores de CHAS, DIS y RAD. Elaboración propia.

La comparación entre las distribuciones obtenidas para la estimación de la media esperada de MEDV en la (Figura **??**) y las correspondientes distribuciones predictivas (Figura 4.2) muestran coherencia en los resultados del modelo. En ambos casos se mantienen los mismos patrones de desplazamiento: los valores de `CHAS = 1` se asocian a mayores niveles de `MEDV`, mientras que el incremento de `log(DIS)` y `RAD` se relaciona con una disminución en el valor medio de las viviendas.

La principal diferencia entre ambas representaciones esta en la dispersión de las distribuciones. Las predicciones presentan colas más amplias y una variabilidad mayor, al incorporar tanto la incertidumbre de los parámetros del modelo como el error aleatorio inherente al proceso de predicción. Por el contrario, las distribuciones de la media esperada reflejan únicamente la incertidumbre asociada a la estimación de los coeficientes. Esta diferencia en la amplitud es consistente con la teoría bayesiana, donde las predicciones contienen una mayor incertidumbre.

En conjunto, ambas aproximaciones ofrecen resultados coherentes: el modelo bayesiano reproduce correctamente la relación entre las variables de localización y el valor medio de las viviendas, mostrando un comportamiento estable y razonable tanto en la estimación del valor esperado como en la predicción de nuevas observaciones.

Capítulo 5

Evaluación del modelo con residuos bayesianos

La evaluación del modelo mediante residuos bayesianos permite analizar hasta qué punto el modelo explica adecuadamente los datos observados. El código mostrado a continuación calcula la probabilidad valores predichos de la variable dependiente MEDV, obtenidos a partir del modelo bayesiano de distribución a posteriori, sean considerados valores atípicos dentro del conjunto de observaciones.

El código empleado ha sido:

```
prob.out=bayesresiduals(fit,theta.sample,2)
par(mfrow=c(1,1))
plot(bhd$MEDV,prob.out)
```

Obteniendo la siguiente gráfica:

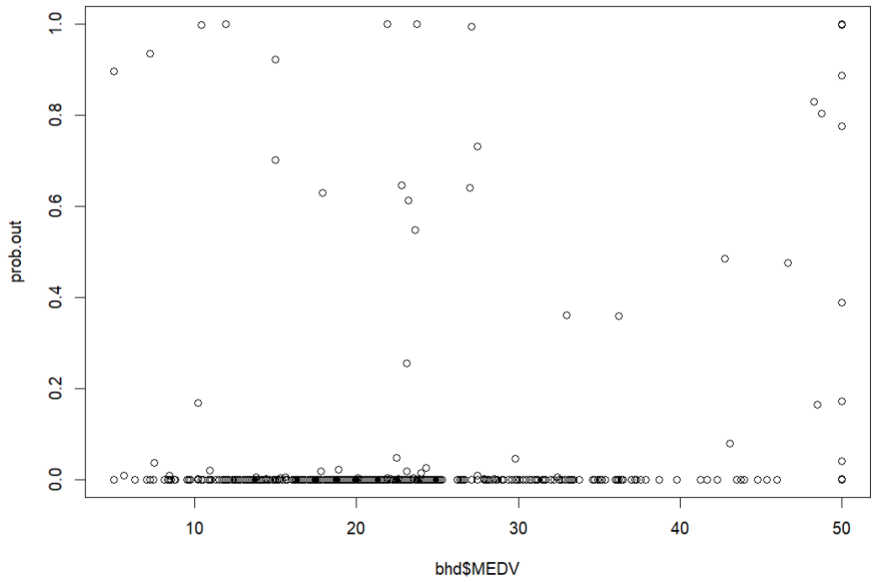

Figura 5.1: Probabilidad a posteriori de observaciones atípicas en función de los valores observados de MEDV. Elaboración propia.

En la Figura 5.1, la mayor parte de las observaciones presentan probabilidades cercanas a cero, lo cual indica que el modelo ajusta adecuadamente los datos en la mayoría de los casos. No obstante, se observan algunas observaciones con probabilidades cercanas a uno, localizadas principalmente en los extremos del

rango de `MEDV`. Esto sugiere la presencia de un pequeño número de valores atípicos o mal explicados por el modelo, posiblemente asociados a situaciones extremas de valor de vivienda, o con cierta atipicidad.

No se observa ningún patrón claro entre los valores de `prob.out` y `MEDV`, lo que sugiere que el modelo comete errores de forma similar en todo el rango de precios. En otras palabras, la variabilidad de los errores se mantiene estable y no parece aumentar ni disminuir según el valor de la vivienda.

En conjunto, el modelo bayesiano parece capturar de manera adecuada la estructura de los datos, aunque con algunas limitaciones en los casos extremos.