Universidad Nacional de Colombia

Facultad de Ciencias
Departamento de Estadística

CASO 3 - ESTADÍSTICA BAYESIANA

EJERCICIOS DEL HOFF: PRÁCTICA DE LOS MÉTODOS BAYESIANOS PARA EL AJUSTE DE MODELOS DE REGRESIÓN LINEAL Y MULTIPLE

Autor:

Cesar Augusto Prieto S.
ceprieto@unal.edu.co

March 6, 2025

Contents

List of Tables

1 Hoff. Ejercicio 9.2

El presente ejercicio tiene como objetivo principal modelar la distribución condicional del nivel de glucosa en plasma sanguíneo (glu) en función de otras variables relacionadas con la salud de una población de mujeres.

Para ello, se trabajará con el conjunto de datos azdiabetes.dat el cual contiene información sobre 532 mujeres y diversas variables biométricas que pueden estar relacionadas con el nivel de glucosa. La variable diabetes está presente en la base, pero se excluye del modelo como variable explicativa. A continuación, se describen las variables consideradas en el análisis:

- **glu:** Nivel de glucosa en plasma sanguíneo.
- **bp:** Presión sanguínea.
- **skin:** Grosor del pliegue cutáneo en tríceps.
- **bmi:** Índice de masa corporal.
- **age:** Edad.
- **npreg:** Número de embarazos.
- **ped:** Probabilidad de desarrollar diabetes en función del historial familiar y factores genéticos.

1.1 Solución 9.2A

Empleando un enfoque bayesiano en el cual se utiliza la previa g con g=n, ν_0=2 y σ_0^2=1, se ajustará el modelo correspondiente para las variables trabajadas y luego se aplicará un procedimiento de selección y promediado de modelos.

Para mostrar un poco de manera gráfica las relaciones que pueden haber entre la variable glucosa y las demás covariables se observan los siguientes diagramas.

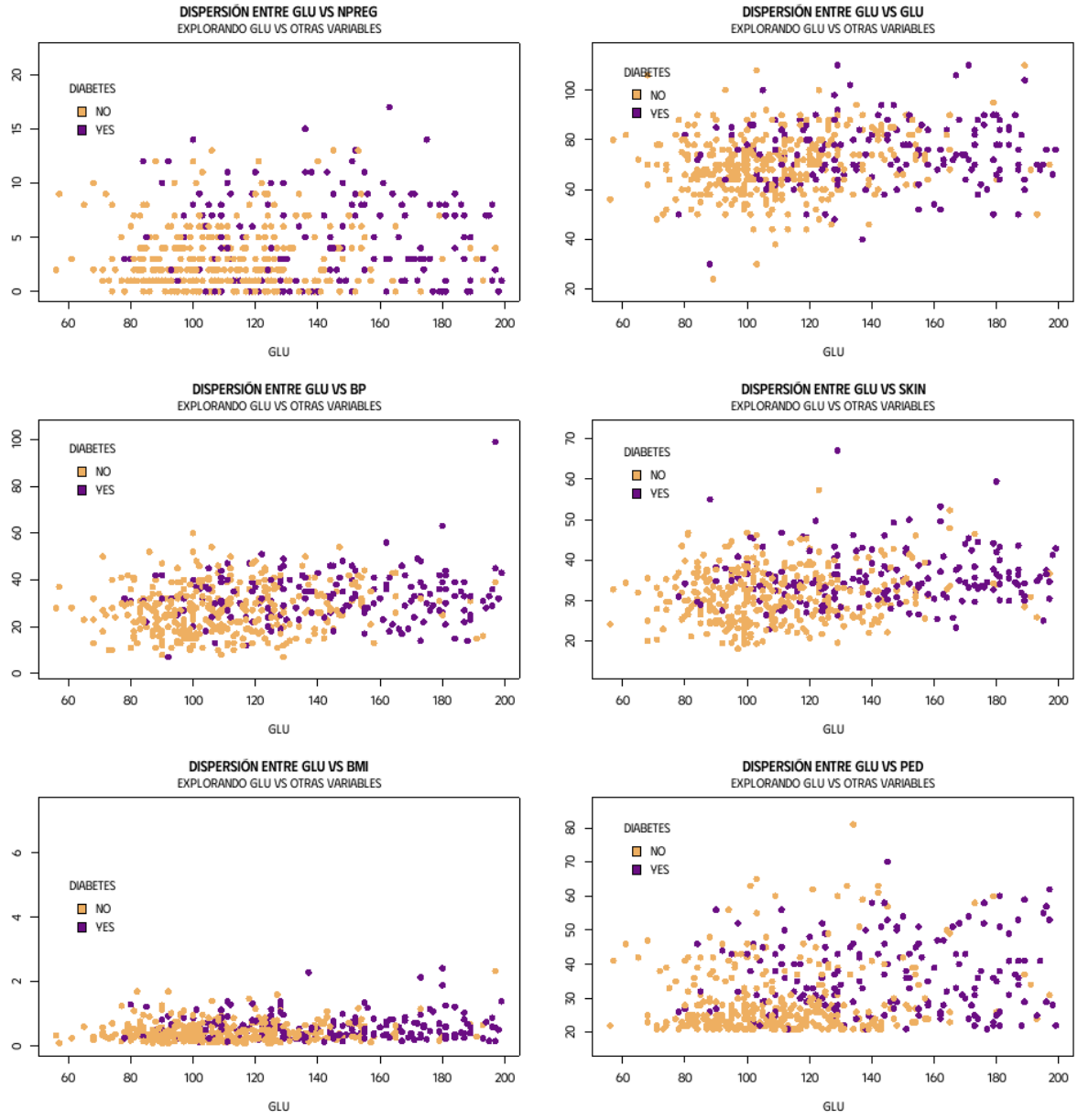

Figure 1

De aquí se tiene el indicio gracias a los gráficos, de que parece haber relación entre la glucosa y variables como por ejemplo, (bmi), (ped) y (age). Ahora, se ajusta el modelo correspondiente.

La siguiente es la tabla de los Betas estimados con sus correspondientes intervalos.

	Beta	2.5%	Mean	97.5%
1	Intercepto	35.140	52.230	69.164
2	Beta_npreg	-1.631	-0.656	0.315
3	Beta_bp	-0.017	0.205	0.431
4	Beta_skin	-0.118	0.193	0.503
5	Beta_bmi	0.150	0.643	1.132
6	Beta_ped	3.262	10.510	17.842
7	Beta_age	0.453	0.765	1.079

Table 1: Intervalos para los parámetros del modelo de regresión lineal

De acá se puede ver gracias a los intervalos que las variables que parecen relacionarse con los niveles de glucosa parecen ser (bmi), (ped) y (age).

De igual forma, también se ajusta un modelo extra de regresión convencional usando la funcion lm de R con el objetivo de comprobar los resultados obtenidos, de aquí se ve que en efecto estas variables anteriormente dichas si guardan relación con la variable glucosa.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercepto	52.305	8.602	6.080	0.000
npreg	-0.657	0.491	-1.338	0.181
bp	0.205	0.113	1.811	0.071
skin	0.193	0.157	1.226	0.221
bmi	0.644	0.247	2.610	0.009
ped	10.548	3.675	2.870	0.004
age	0.767	0.159	4.831	0.000

Table 2: Resumen del modelo de regresión lineal de Prueba (lm)

1.2 Solución 9.2B

Esta tabla que sigue representa la probabilidad de inclusión y los límites del intervalo para cada una de las variables tratadas o dicho de otro modo, los betas en el modelo.

	parámetro	prob	lím inf	lím sup
1	intercept	1.000	42.922	77.001
2	npreg	0.100	-1.002	0.000
3	bp	0.167	0.000	0.315
4	skin	0.082	0.000	0.319
5	bmi	0.988	0.445	1.332
6	ped	0.669	0.000	17.150
7	age	1.000	0.479	1.013

Table 3: Probabilidad de inclusión e intervalos para los betas estimados

De aquí, se puede decir que la probabilidad de β_j dado y, será cercana a 1 para el intercepto, la edad (age), el índice de masa corporal (BMI) y también se podría decir que alta para la variable (ped) con 0.67, a su vez, se comenta que será baja con valores de 0.1 aproximadamente para la variable número de embarazos (npreg) y el grosor de la piel (skin), y por último de 0.17 para la presión arterial (bp). Si se compara esta parte con los resultados obtenidos en a), no solo se podrán ver ciertas concordancias sino que a su vez, que para un gran número de variables la longitud de los intervalos tiende a ser más corta, lo cual tal vez se podría traducir en brindar más precisión y significancia en las estimaciones.

De manera adicional, el gráfico representa la probabilidad de inclusión de cada beta dentro de un modelo final.

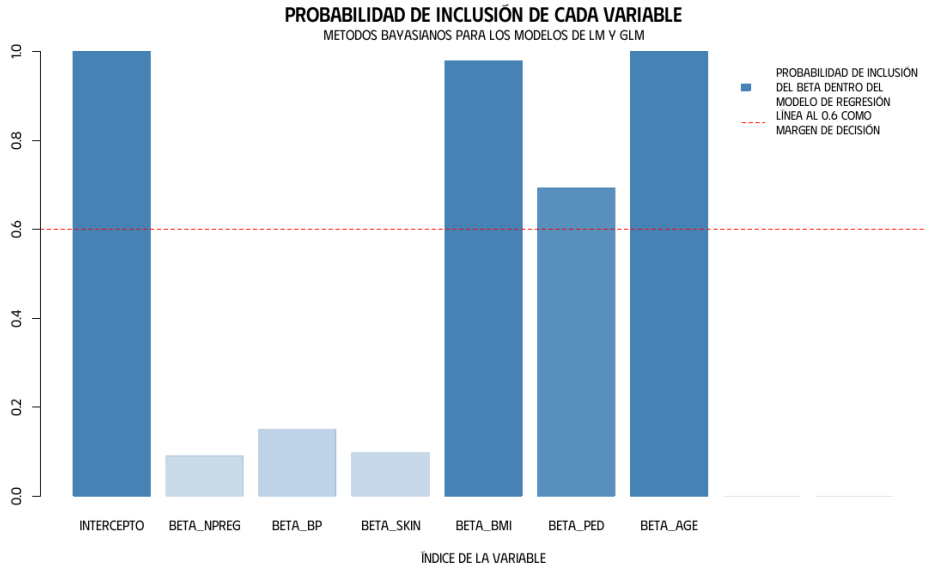

Figure 2

2 Hoff. Ejercicio 9.3

2.1 Solución 9.3A

Luego de ajustar el modelo correspondiente, se muestra como sigue el resultado de la regresión bayesiana utilizando la previa g.

(Falta la primera tabla con los valores de la g prioir!!!, solo esta la tabla de OLS)

	param	beta_hat	pvalue	ci_lower	ci_upper	ci_length
1	intercept	-0.000	0.995	-0.161	0.161	0.322
2	M	0.287	0.043	0.010	0.563	0.554
3	So	-0.000	1.000	-0.375	0.375	0.751
4	Ed	0.545	0.005	0.178	0.911	0.733
5	Po1	1.472	0.081	-0.194	3.137	3.331
6	Po2	-0.782	0.366	-2.519	0.955	3.474
7	LF	-0.066	0.670	-0.379	0.247	0.626
8	M.F	0.131	0.404	-0.185	0.448	0.633
9	Pop	-0.070	0.584	-0.329	0.189	0.518
10	NW	0.109	0.531	-0.242	0.460	0.701
11	U1	-0.271	0.179	-0.672	0.130	0.802
12	U2	0.369	0.049	0.002	0.736	0.734
13	GDP	0.238	0.365	-0.290	0.766	1.056
14	Ineq	0.726	0.004	0.249	1.204	0.955
15	Prob	-0.285	0.041	-0.558	-0.012	0.546
16	Time	-0.062	0.642	-0.329	0.206	0.534

Table 4: OLS Coeficientes e intervalos

De aquí en primera instancia al ver los p valores, se diría que con una significancia de 5%se deben incluir en el modelo las variables (M), (Ed), (U_2), $(Ineq)$ y $(Prob)$.

A continuación se muestra un gráfico comparativo.

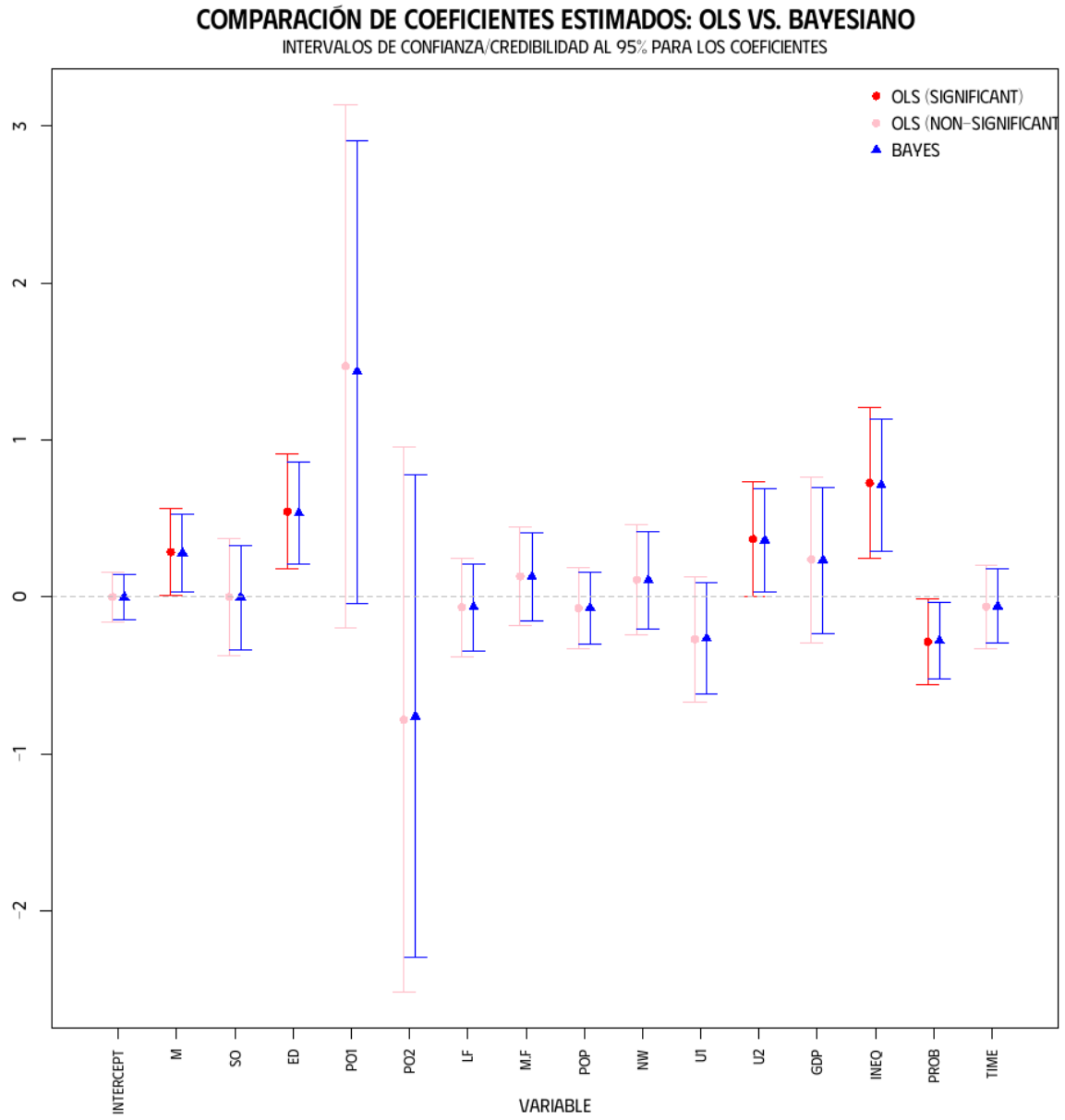

Figure 3

Analizando lo anteriormente obtenido, se tiene que las estimaciones puntuales y los intervalos de credibilidad/confianza del 95% son muy similares entre la regresión bayesiana y la estimación por mínimos cuadrados. La longitud de los intervalos tiende a ser más corta en la regresión bayesiana que en la estimación por mínimos cuadrados.

Al observar los resultados de la regresión bayesiana, los intervalos de credibilidad del 95% para los coeficientes de M, Ed, Po1, U2, Ineq y Prob no incluyen el 0, por lo que estas

variables parecen ser fuertemente predictivas de las tasas de criminalidad.

Por otro lado, los resultados de la estimación por mínimos cuadrados muestran que los valores p de M, Ed, U2 e Ineq son menores a 0.05, lo que indica que estas variables parecen ser fuertemente predictivas de las tasas de criminalidad.

Usando estos criterios, la diferencia entre ambos métodos es que en la regresión bayesiana, Po1 y Prob son predictivos de las tasas de criminalidad, mientras que en la estimación por mínimos cuadrados no lo son.

2.2 Solución 9.3B

A continuación se muestran un par de gráficos con los valores predichos para cada modelo.

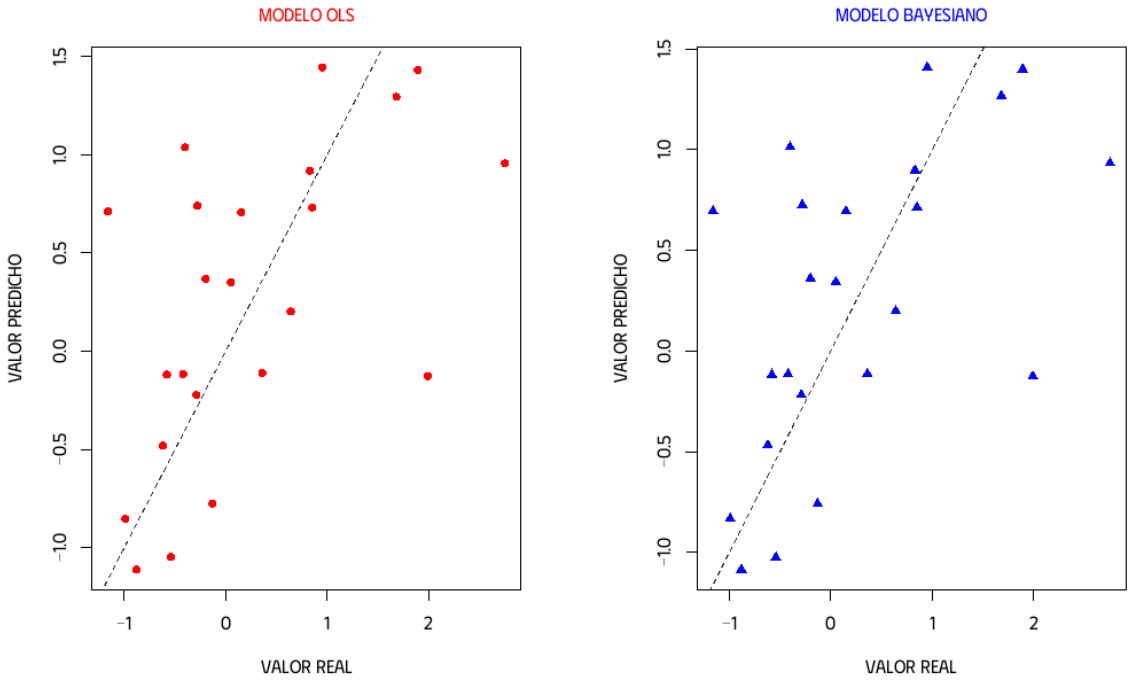

Figure 4

A partir de los resultados anteriores, la regresión por mínimos cuadrados y la regresión bayesiana parecen generar predicciones muy similares para los datos de prueba. En esta ocasión, la regresión bayesiana predice mejor que la regresión por mínimos cuadrados, pero la diferencia es muy pequeña.

2.3 Solución 9.3C

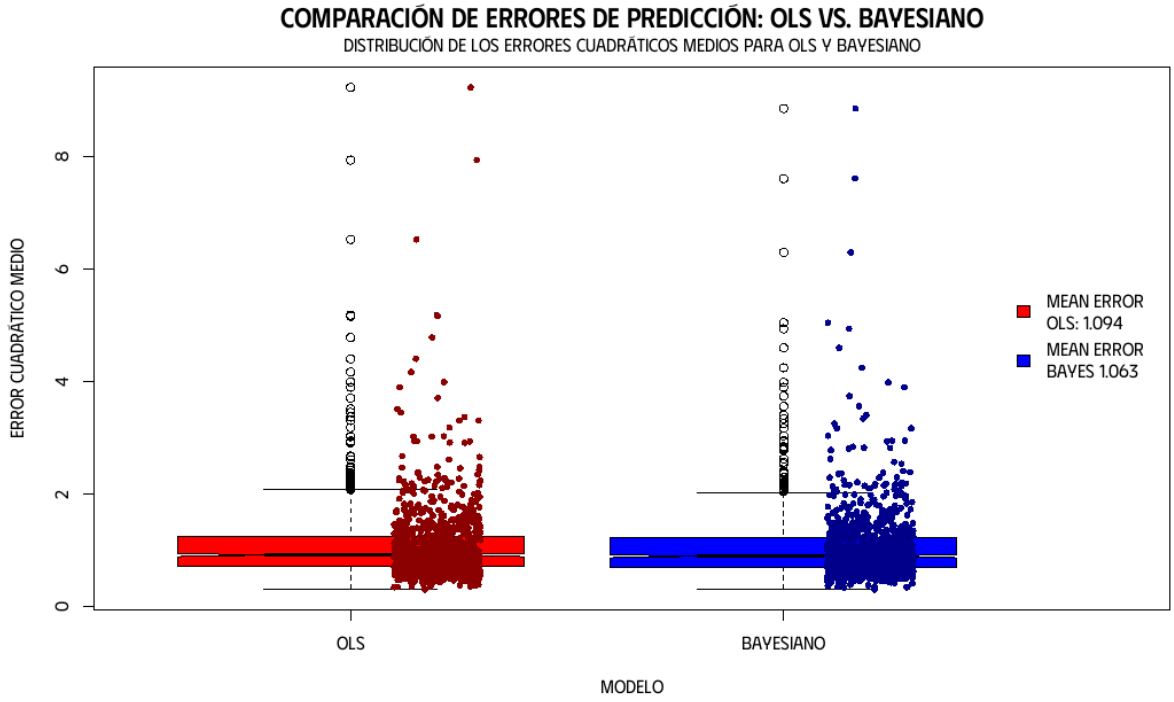

Figure 5

En el gráfico presentado, se comparan los errores de predicción del modelo de Regresión Lineal Ordinaria (OLS) y del modelo Bayesiano. Se observa que ambos modelos exhiben un comportamiento similar en términos de precisión de las predicciones. Sin embargo, al analizar el error medio (MEAN ERROR), se evidencia que el modelo Bayesiano logra un desempeño ligeramente superior, al obtener un error medio menor en comparación con el modelo OLS. Esta diferencia, aunque sutil, sugiere que la aproximación Bayesiana ofrece una ventaja en la reducción del error de predicción.

3 Hoff. Ejercicio 10.3

3.1 Solución 10.3A

```
    lmfit <- lm(height ~ time + pH, data = df);summary(lmfit)

Call:
lm(formula = height ~ time + pH, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1225 -0.4307 -0.1101  0.4574  1.5301

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.2087     0.5891   12.24 7.45e-10 ***
time          3.9910     0.3280   12.17 8.14e-10 ***
pH            0.5778     0.1204    4.80 0.000167 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7334 on 17 degrees of freedom
Multiple R-squared:  0.9096, Adjusted R-squared:  0.899
F-statistic: 85.55 on 2 and 17 DF,  p-value: 1.339e-09
```

El modelo de regresión lineal ajustado para predecir la variable height en función de las variables time y pH mostró un ajuste significativo y explicativo. Los coeficientes estimados para ambas variables predictoras fueron estadísticamente significativos, con valores de p inferiores a 0.001. El coeficiente de la variable time fue estimado en 3.9910, lo que indica que, manteniendo constante el pH, un aumento en el tiempo se asocia con un incremento significativo en la altura. Por otro lado, el coeficiente de la variable pH fue estimado en 0.5778, sugiriendo que, manteniendo constante el tiempo, un aumento en el pH también está relacionado con un incremento en la altura, aunque en menor magnitud.

El modelo presentó un alto poder explicativo, con un coeficiente de determinación ajustado (R^2 ajustado) de 0.899, lo que indica que aproximadamente el 89.9% de la variabilidad en la altura puede ser explicada por las variables time y pH. Además, el error estándar residual fue de 0.7334, lo que refleja una buena precisión en las predicciones del modelo.

En conclusión, el modelo de regresión lineal ajustado fue considerado adecuado para

describir la relación entre la altura y las variables predictoras time y pH, demostrando que ambas variables tienen un impacto significativo en la respuesta. Estos resultados sugieren que el modelo puede ser utilizado para predecir la altura en función de estas variables, siempre dentro del rango de los datos analizados.

3.2 Solución 10.3B

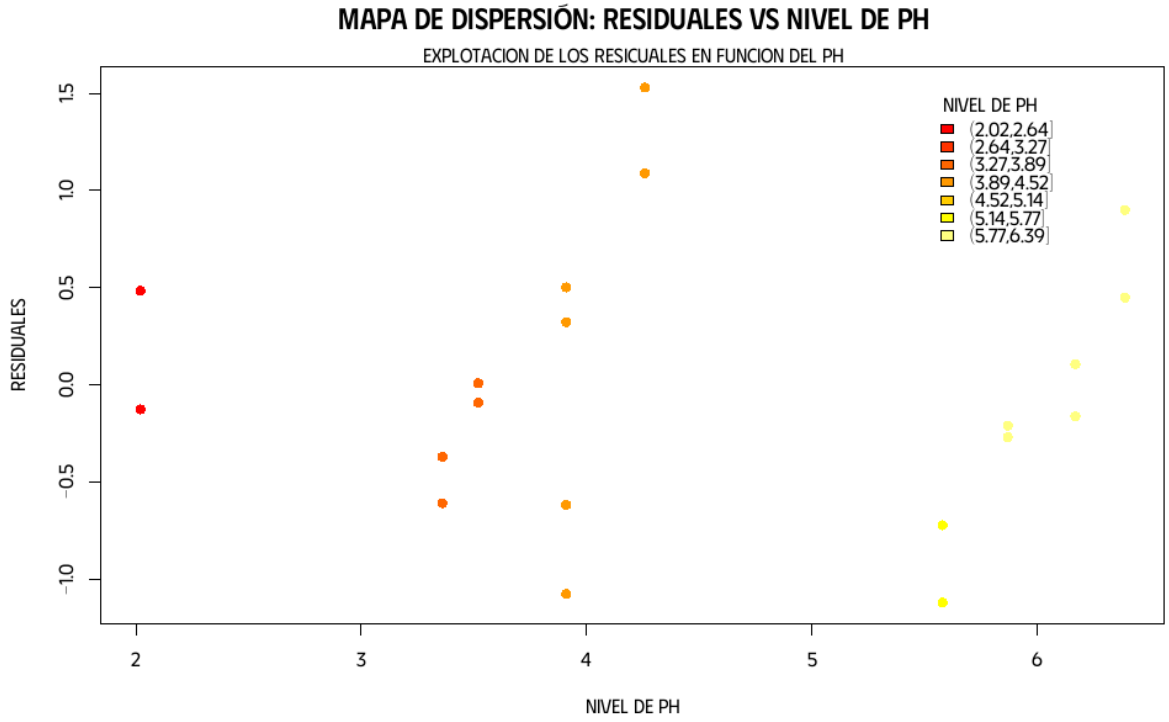

Figure 6

El diagrama de dispersión de los residuos muestra que los valores correspondientes a la misma planta de tomate están correlacionados, lo que indica que los errores no son independientes. Esto viola el supuesto fundamental del modelo de mínimos cuadrados ordinarios (MCO) de que los errores deben estar distribuidos de forma independiente e idéntica (i.i.d.). Como consecuencia, los coeficientes de regresión y sus errores estándar pueden estar sesgados, afectando la validez de las inferencias.

3.3 Solución 10.3C

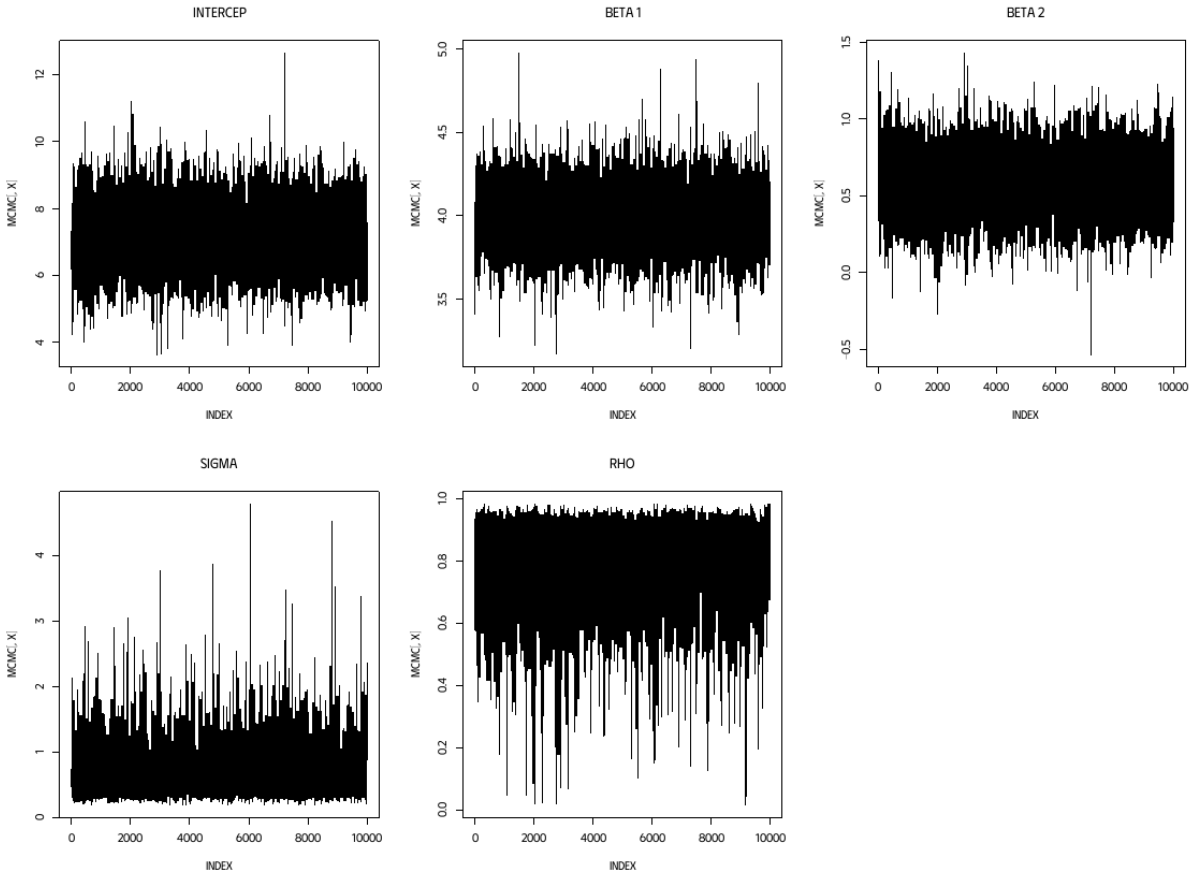

Figure 7

La figura observada contiene las cadenas de convergencia de los parámetros del modelo de regresión ajustado mediante el algoritmo de Metropolis, en las cuales se puede observar cómo cada uno de los parámetros convergió a lo largo de la cadena.

3.4 Solución 10.3D

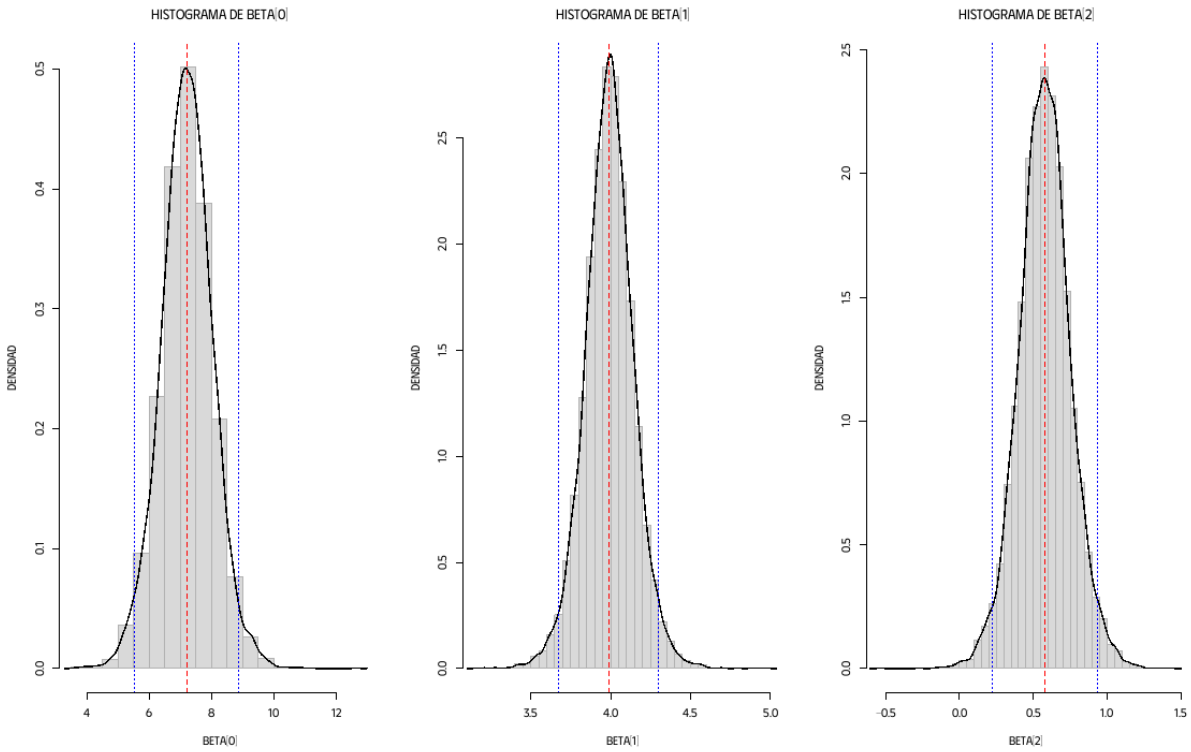

Figure 8

	param	mean	lower	upper
1	Intercep	7.210	5.519	8.871
2	beta 1	3.992	3.677	4.302
3	beta 2	0.577	0.223	0.937
4	sigma	0.649	0.289	1.456
5	rho	0.791	0.440	0.950

Table 5: Estadísticas de los parámetros del modelo

El modelo de regresión lineal ajustado para predecir la variable ‘height‘ en función de ‘time‘ y ‘pH‘ demostró ser altamente significativo y explicativo. Los coeficientes estimados para ‘time‘ (3.992) y ‘pH‘ (0.577) fueron estadísticamente significativos, confirmando su impacto positivo en la altura. El intercepto (7.210) y los intervalos de confianza de los parámetros, junto con la convergencia observada en los gráficos, respaldan la robustez del modelo.

Los resultados son consistentes con los obtenidos en la regresión lineal clásica, lo que valida su confiabilidad. En conclusión, el modelo es adecuado para predecir la altura en función de ‘time‘ y ‘pH‘, aunque podrían explorarse mejoras incorporando otras variables o analizando estructuras residuales.

4 Hoff. Ejercicio 11.4