Regresión Avanzada

Joaquín Álvarez Galnares

Febrero 2023

Abstract

En este segundo reporte presentamos un análisis e interpretación de resultados correspondientes al ejercicio 4 de las notas 4 de clase en las cuales se cubre el tema de modelos lineales generalizados con un enfoque bayesiano. El código correspondiente para reproducir los resultados se encuentra **aquí**.

Contexto

Consideramos un modelo de regresión lineal múltiple

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \epsilon_i, \quad i = 1, \ldots, n$$

donde Y_i es el salario anual (en miles de dólares) del i-ésimo trabajador de una compañía de nivel medio y alto, $X_{1,i}$ es el índice de calidad de su trabajo, $X_{2,i}$ su número de años de experiencia y $X_{3,i}$ el índice de éxito en sus publicaciones. Además supondremos que $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau)$. Consideremos las siguientes distribuciones iniciales no informativas:

$$\alpha, \beta_j \sim \mathcal{N}(0, 0.001), \quad j = 1, 2, 3$$

$$\tau \sim Gamma(0.001, 0.001)$$

Inferencia bayesiana para los parámetros vía MCMC en JAGS

Se simuló una cadena con $10,000$ iteraciones y con un periodo de calentamiento de 1000 iteraciones dando lugar a una cadena de $9,000$ iteraciones y se usó un adelgazamiento de una unidad.

Notamos que la línea de los promedios ergódicos se estabiliza alrededor de la iteración 1,000. La cadena tiene niveles de autocorrelación bajos y la traza recorre "rápidamente" el intervalo $[10, 25]$. Es relevante señalar que la interpretabilidad del intercepto en este modelo de regresión lineal podría depender

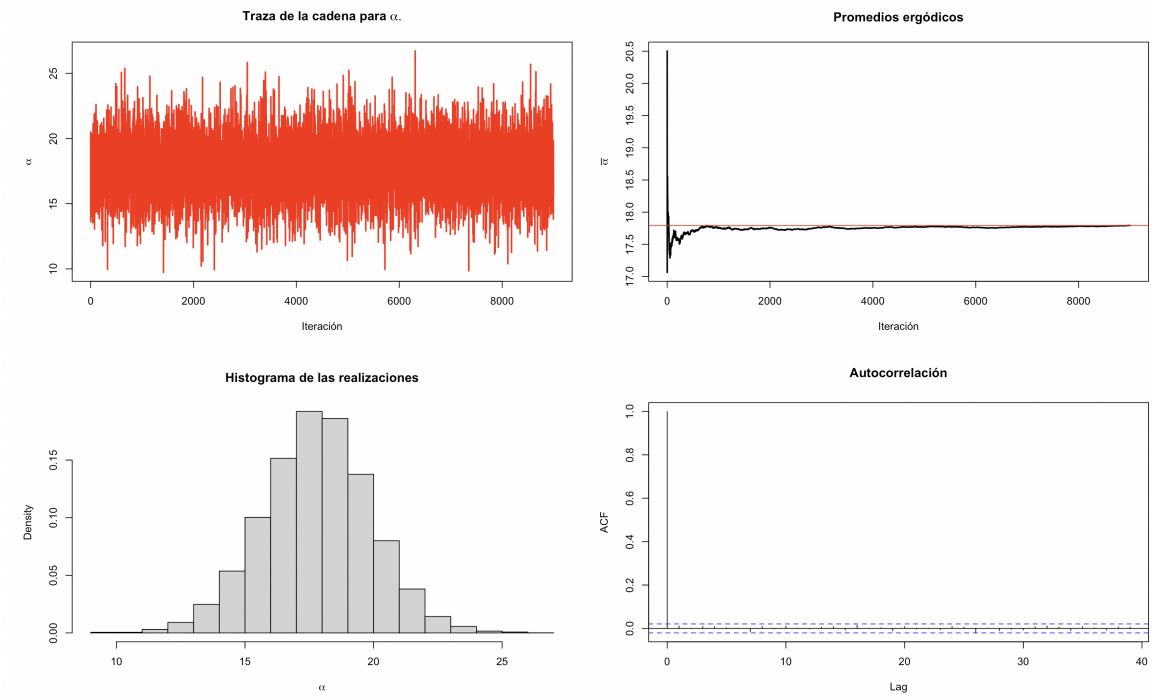

Figure 1: Cadena de α

del contexto. Por ejemplo quizá tenga sentido pensar en que elintercepto nos indica el salario anual promedio de una persona que no tiene experiencia laboral, tiene una calificación de 0 en el índice de la calidad de su calidad de trabajo y tiene una calificación de cero en sus publicaciones. Pero por ejemplo, posiblemente el índice de calidad de trabajon de una persona núnca pueda tomar el valor 0. Sin embargo, es razonable suponer que es interpretable. Y en dicho caso, se interpreta como el valor que toma la variable de respeusta cuando todas las variables explicativas valen cero.

Al observar la traza de la cadena para β_1 notamos que en pocas ocasiones la cadena toma valores negativos. Los promedios ergódicos dejan de tener oscilaciones alrededor de la iteración 1000. Es muy interesante apreciar la convergencia que garantiza el Teorema Ergódico. La línea horizontal naranja en la trayectoria de los promedios ergódicos corresponde al límite, garantizao por el Teorema Ergódico, que en este caso, a nivel finto (9 mil iteraciones), lo estimamos con la media de la cadena. La gráfica inferior derecha de la figura 2 sugiere bajos niveles de correlación en la cadena.

La forma de leer la figura 3 es análoga a la que se presentó para leer la figura 2.

En la figura 4 se obtienen patrones relativamente similares a los de las figuras previas. La distribución posterior tiene una forma acampanada simétrica con colas ligeras.

Respecto a la cadena de τ, podemos notar que la velocidad de convergencia es relativamente similar a la convergencia a la distribución límite que observamos en otros parámetros. Algo diferente es la distribución límite. En la figura 6 presentamos un histograma más amplio para que se aprecie que la posterior de $\tau|(X, \underline{y})$ presenta una ligera asimetría con un sesgo a la derecha(una cola más pesada a la deracha.)

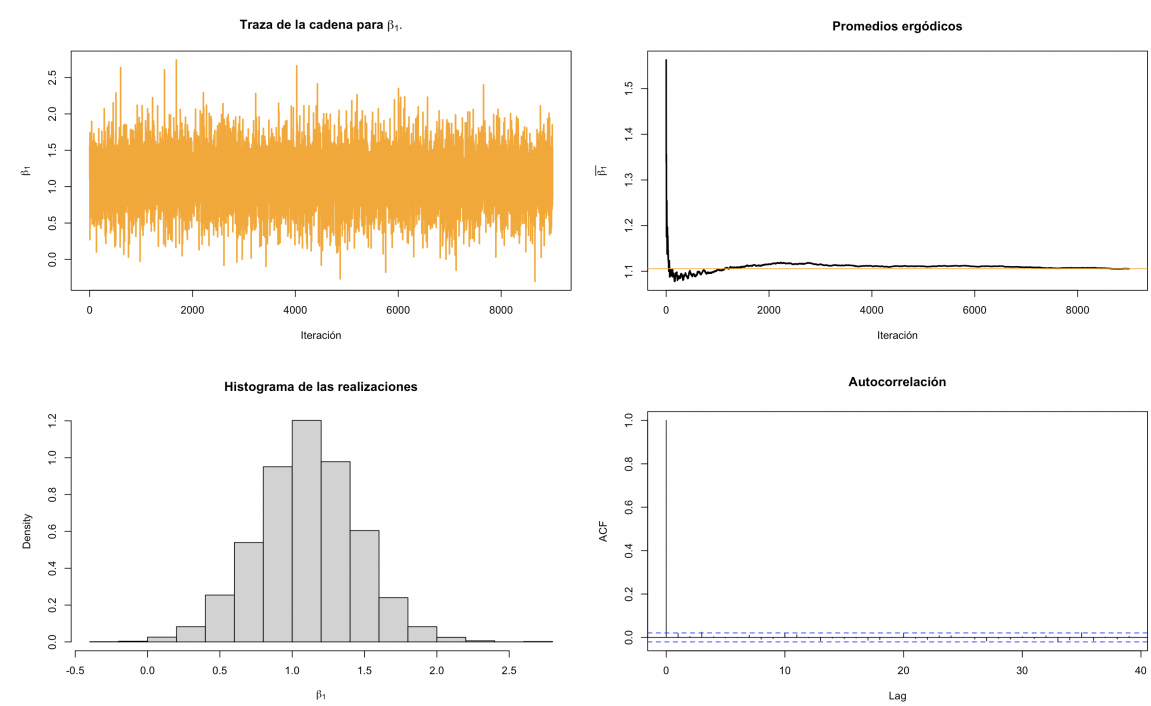

Figure 2: Cadena de β_1

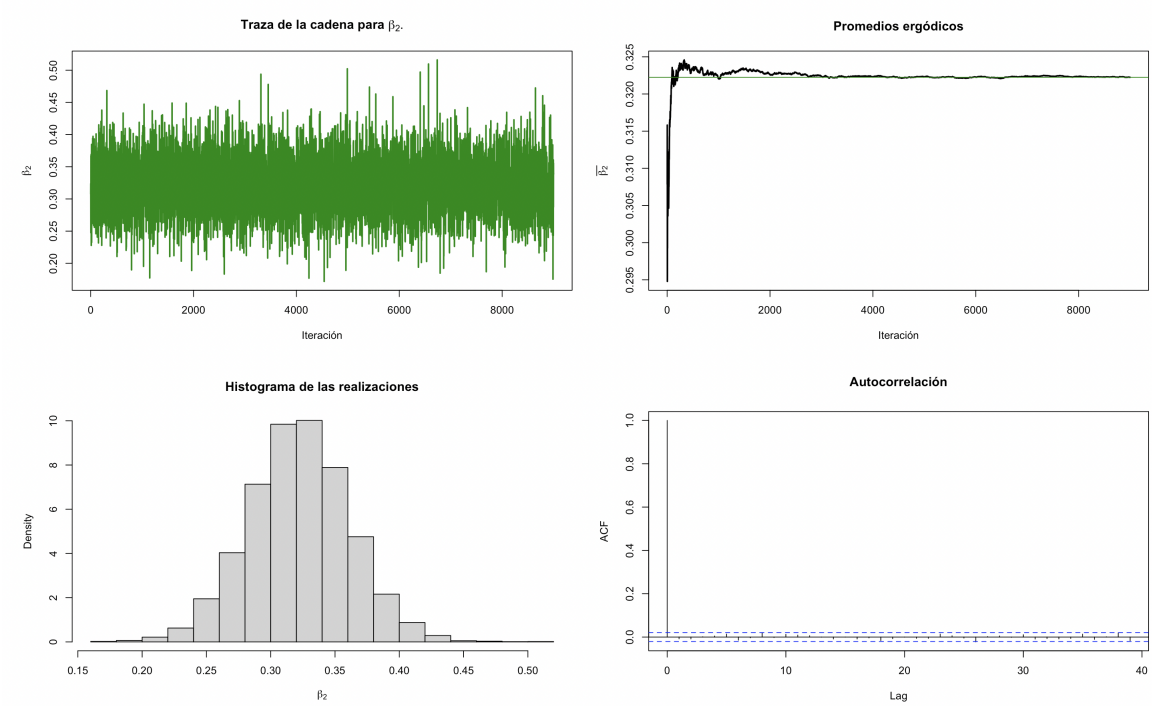

Figure 3: Cadena de β_2

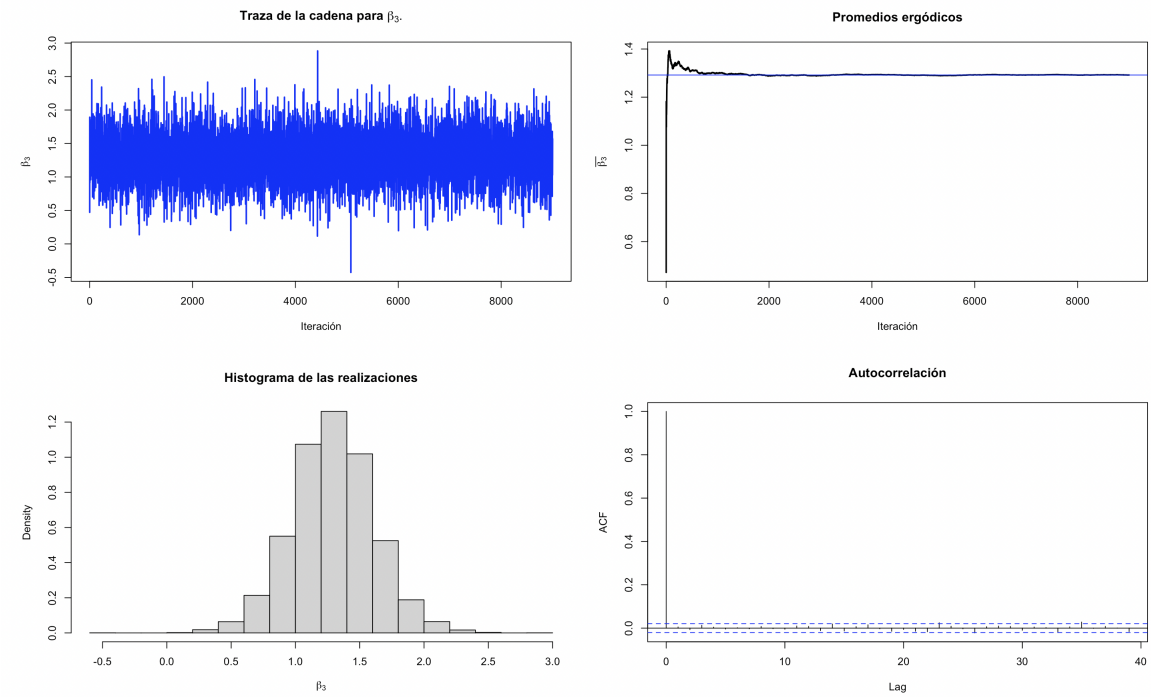

Figure 4: Cadena de β_3

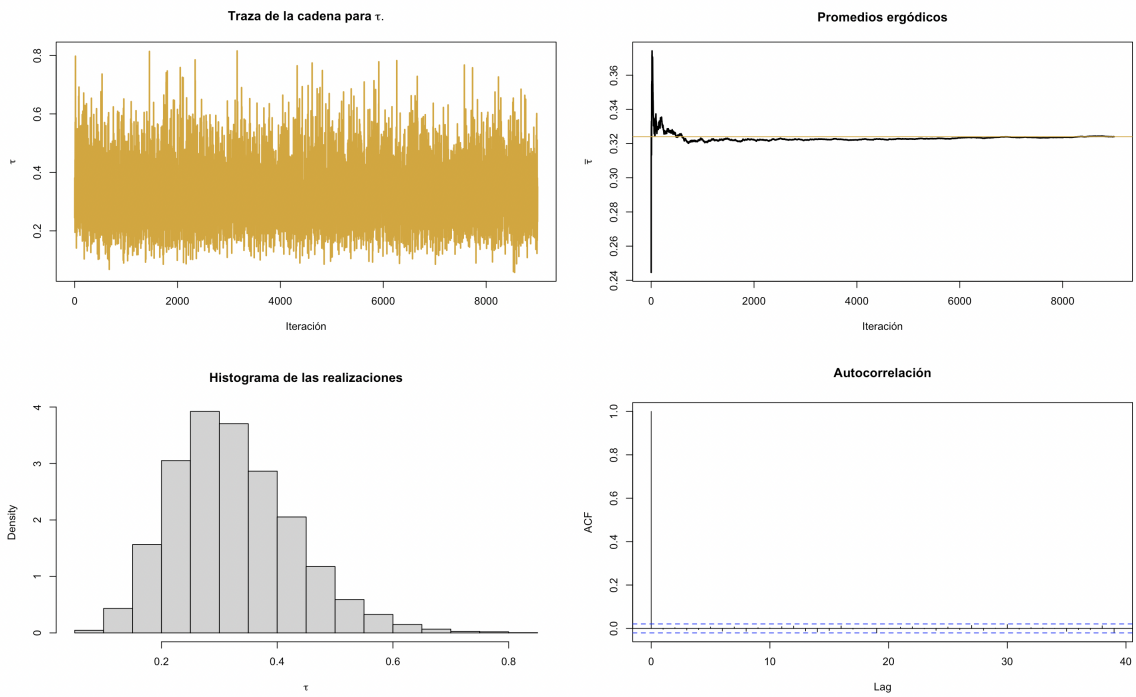

Figure 5: Cadena de τ

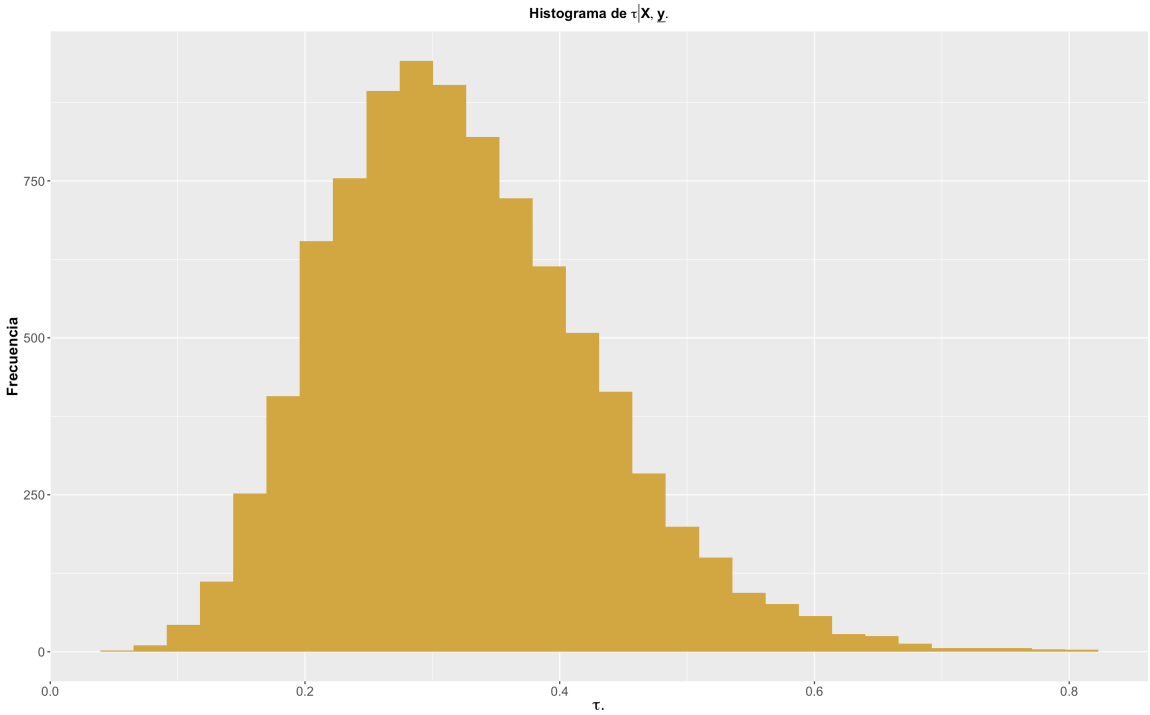

Figure 6: Histograma que muestra la forma de la distribución posterior de τ

Análisis de los parámetros que afectan a la media de la variable respuesta condicionalmente: $\alpha, \beta_1, \beta_2, \beta_3$

Notamos que todas las distribuciones tienen una forma acampanada y simétrica de colas no muy pesadas, por lo cual se ajustan bien a una distribución normal (figura 7). Los 4 coeficientes toman valores positivos. Solo hay poca frecuencia de observaciones negativas para β_1, pero son desperciables, como veremos más adelante al reportar los cuantiles $q0.025$ y $q0.975$ de cada distribución posterior.

En la figura 8 podemos apreciar que si formamos un diagrama de dispersión para los coeficientes en pares, estos presentan forma de elipses, lo cual sugiere que las distribuciones conjuntas de los coeficientes, por ejemplo la conjunta de la posterior (α, β_1) tiene una distribución normal bivariada, o, como señala la teoría que vimos en las notas de clase, una t de student bivariada. La pendiente del eje principal de las elipses también sugiere una correlación negativa entre algunos coeficientes. Elipses (nubes de puntos) más horizontales de puntos sugieren relaciones lineales más debiles (correlaciones cercanas a cero). Se obtuvo una correlación de -.63 entre α y β_3, lo cual sugiere una fuerte relación lineal. Visualmente la podemos apreciaren el diagrama superior de la esquina derecha. Mientras que, α y β_2 tuvieron una correlación cercana a cero.

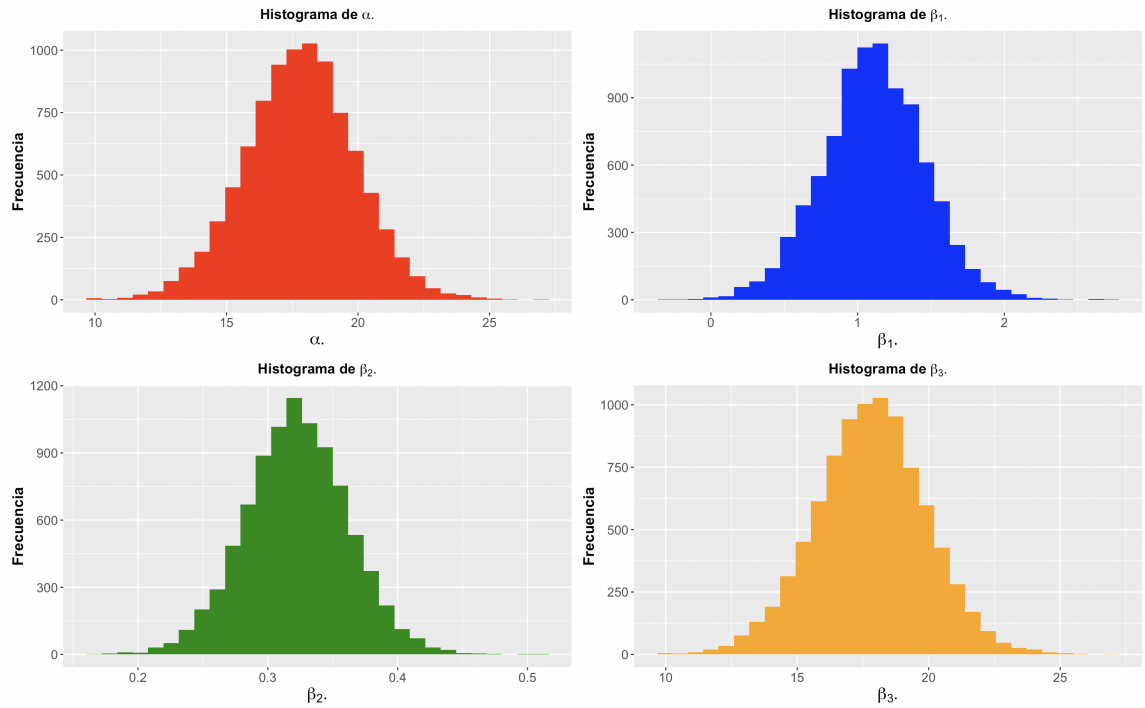

Figure 7: Histograma de la posterior marginal de cada parámetro que afecta a la media condicional del ingreso anual de los empleados

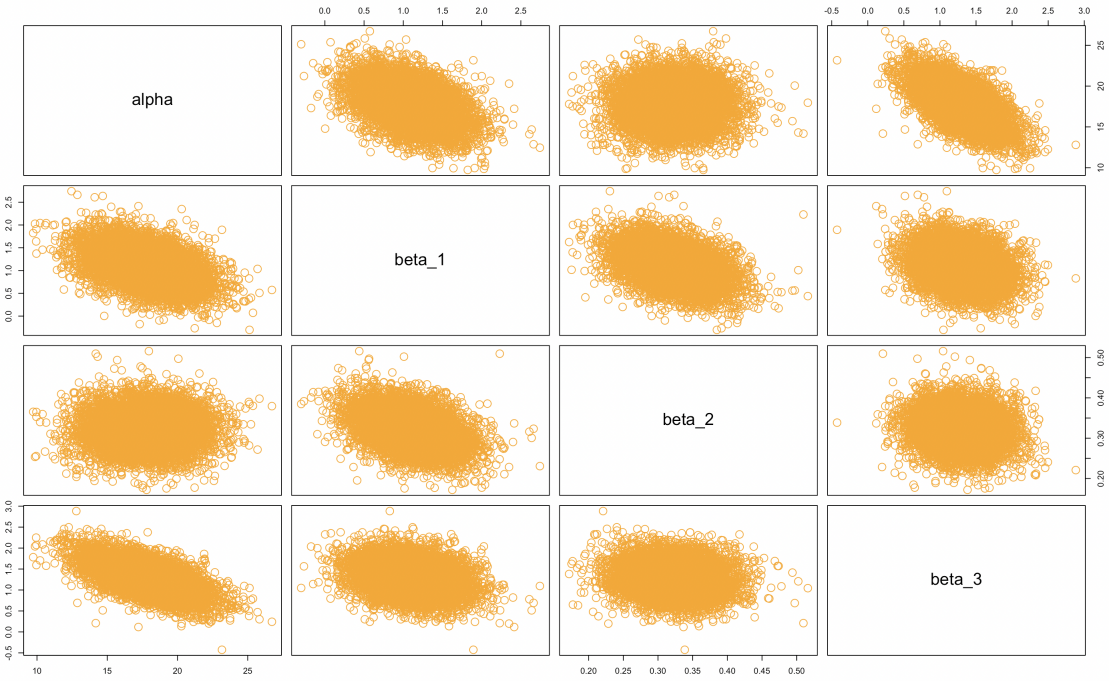

Figure 8: Correlación de los coeficientes que afectan a la media condicional de la varaiable respuesta

Resúmen de inferencia bayesiana para los parámetros del modelo de regresión lineal múltiple

Parámetro	Estimación puntual	Intervalo al 95% de confianza
α	17.7928	(13.5859, 21.896)
β_1	1.1057	(0.4117, 1.7846)
β_2	0.3222	(0.0396, 0.4004)
β_3	1.2920	(0.31979, 1.94194)
τ	0.3240	(0.1050, 0.5614)

La estimación puntual utilizada es la media estimada para cada distribución posterior marginal utilizando la dadena que simulamos. Dicho estimador puntual está asociado a la minimización de una pérdida cuadrática. Notamos que todos los coeficientes que afectan a la media condicional del modelo tienen un valor positivo en su estimación puntual. Además, en ningún caso se incluye al cero en el intervalo de confianza, por lo cual esto puede tomarse como evidencia estadística de que todas las variables explicativas que se están utilizando son significativas al 95%. Es importante el signo de las betas. En particular, por que al ser positivas las estimaciones puntuales, esto nos dice que todas tienen un efecto positivo sobre la media de la varaible de respuesta, de manera que dejando el resto de las variables fijas, incrementar una unidad en la cualquier variable explicativa, incrementa el promedio de la variable de respuesta(el ingreso de los empleados). Así por ejemplo, si se incrementa en un año los años de experiencia del empleado, entonces, en promedio, el empleado percibirá un mayor salario. Los signos también serán relevantes para analizar las predicciones finales usando nuevos datos, como veremos más adelante: pues si un individuo tiene todos sus atributos (variables explicativas) que otro individuo, sería razonable que quien tenga valores más grandes en las variables explicativas tenga asociada una mayor predicción puntual en su ingreso.

Consistente con lo que vimos en clase, las medias quedaron muy cercanas a las que se obtienen con

$$\underline{\hat{\beta}} = \begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 17.8469 \\ 1.1031 \\ 0.3215 \\ 1.2889 \end{bmatrix}$$

Comparando las estimaciones puntuales, todas son muy cercanas a lo que se obtiene con la estimación de mínimos cuadrados ordinarios (OLS), que como vimos en clase, no es mera coincidencia. Hay conexiones entre ese estimador y la media de la posterior de los coeficientes del modelo de regresión lineal.

Predicciones sobre nuevas observaciones

Consideramos predicciones de los salarios para 3 nuevos empleados con los siguientes datos en las variables explicativas:

$$\underline{X}_{F_1} = \begin{bmatrix} 5.4 \\ 17 \\ 6.0 \end{bmatrix}, \quad \underline{X}_{F_2} = \begin{bmatrix} 6.2 \\ 12 \\ 5.8 \end{bmatrix}, \quad \underline{X}_{F_3} = \begin{bmatrix} 6.4 \\ 21 \\ 6.1 \end{bmatrix}$$

Es relevante destacar algo intuitivo que refleja que la distribución predictiva tiene sentido. Por ejemplo, notamos que el individuo 3 con vector de covaraibles $\underline{x}_{F_3}$ tiene valores más grandes en sus tres atributos en comparación a los otros dos individuos. Y, una vez que por la inferencia tenemos que las tres variables explicativas tienen un efecto positivo condicionalmente en la media de la variable de respuesta, entonces es lógico que la distribución final de $Y_{F_3}|(\underline{x}_{F_3}, X, \underline{y})$ tenga una densidad mayor para valores más grandes de la variable de respuesta, en comparación a$Y_{F_1}|(\underline{x}_{F_1}, X, \underline{y})$ y a la densidad de $Y_{F_2}|(\underline{x}_{F_2}, X, \underline{y})$. Así pues, por esta razón, lo coherente es que la predicción puntual para Y_{F_3} sea mayor que la de Y_{F_1} y que la de Y_{F_2}.

Usando el resúmen que reporta JAGS, la media de la distribución predictiva para cada variable, fue de 37.030, 36.025, 39.546, respectivamente, lo cual es coherente con las observaciones respecto a las predicciones puntuales. Aprovechando los cuantiles que reporta el resúmen de JAGS $q_{0.025}$ y $q_{0.975}$ podemos construir intervalos de .95 de probabilidad, de manera que

Variable respuesta	Predicción puntual	Intervalo de 0.95 de Probabilidad
Y_{F_1}	37.03	(33.218, 40.759)
Y_{F_2}	36.02	(32.041, 39.984)
Y_{F_3}	39.54	(35.655, 43.565)

La predicción puntual utilizada es la media estimada para cada distribución predictiva final, la cual corresponde a la minimización de una pérdida cuadrática. Es relevante enfatizar que el razonamiento para comparar la predicción puntual solo aplica con la de Y_{F_3} respecto a las otras, puesto que los atributos asociados al tercer empleado son todos mayores que los de los otros empleados. Las tres distibuciones tuvieron una desviación estándar muy cercana, todas cercanas a 1.9. Esto puede apreciarse visualmente si se consideran lsa densisdades de la figura 9.

Medida de ajuste del modelo

La pseudo-R cuadrada obtenida fue de 0.911

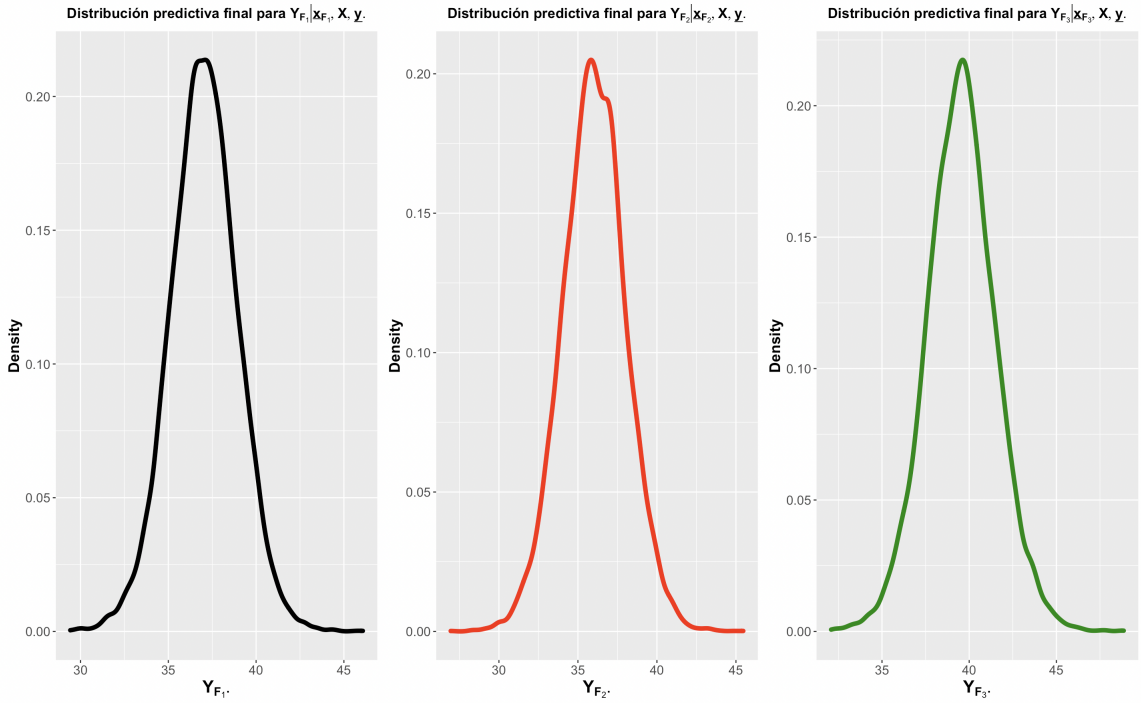

Figure 9: Distribuciones predictivas finales para el salario anual de cada empleado.