AULA ESTADÍSTICA

Serie Monográfica: Introducción a la Regresión Bayesiana

Autor: Jorge Lorenzo

Octubre 2024

Recursos Educativos Abiertos

Introducción a la Regresión Bayesiana

Introducción

La regresión bayesiana es un enfoque estadístico para modelar y analizar la relación entre variables dependientes e independientes usando el teorema de Bayes, es decir, se modela la incertidumbre en los parámetros a través de una distribución de probabilidad conocida (generalmente la distribución normal estándar N(0; 1). En una regresión tradicional, los valores predichos de la variable dependiente, resultan en una combinación lineal de las variables predictoras, y una vez establecidos los coeficientes de la ecuación, estos no pueden ser actualizados con nuevos datos. La regresión bayesiana tiene la ventaja de incorporar nueva información sobre los parámetros del modelo a medida que obtenemos nuevos datos, lo cual le da mayor versatilidad. Este enfoque es preferible en situaciones en la que se tiene información previa relevante sobre los parámetros del modelo que se van a incorporar en el análisis. Generalmente esto se consigue cuando el fenómeno de interés ha sido indagado en profundidad en investigaciones previas. Por defecto, la regresión lineal utiliza la distribución normal estándar como modelo, si bien la regresión bayesiana también lo utiliza, es posible adoptar otros modelos de distribuciones a priori para incluir conocimientos anteriores o resultados de estudios previos, en tal sentido la regresión bayesiana es más flexible. Esto es útil cuando se propone una hipótesis con parámetros conocidos, que determinan una expectativa sobre el comportamiento de los datos que debe ser recogido en el análisis.

En investigación educativa, la regresión bayesiana tiene ventajas sobre la regresión tradicional, pues permite modelizar mejor conjuntos pequeños de datos, producto de un tamaño muestral acotado; en tales situaciones el modelo lineal estándar resulta inestable, variando sus resultados en réplicas sucesivas. El modelo bayesiano en cambio, al incorporar información a priori, es más robusta en situaciones de datos limitados. Por otro lado, el modelo bayesiano permite trabajar sobre la estimación de la incertidumbre completa, al aportar las distribuciones posteriores para los parámetros, lo que permite obtener una descripción más detallada de la incertidumbre.

Cuando el tamaño de la muestra es grande, la regresión bayesiana resulta ventajosa frente a la regresión estándar, en que pueden incluir múltiples niveles de la variable dependiente, anidados en modelos no lineales. En estadística educativa esta es una situación común que se resuelve a través de los modelos multinivel o jerárquicos; la regresión bayesiana permite incorporar modelos no lineales dinámicos, que se ajustan a medida que se incorporan nuevos datos. Al no depender de ecuaciones lineales, la multicolinealidad (variables independientes altamente correlacionadas), no resulta un problema y los coeficientes de regresión son más estables.

Algunos conceptos centrales de la regresión bayesiana que es necesario conocer para su correcta interpretación son los siguientes:

1. **Distribución Priori**: representa nuestras creencias iniciales sobre los parámetros antes de observar los datos. La aplicación del Teorema de Bayes permite cambiar las

estimaciones basadas en estos parámetros iniciales, a medida que obtenemos nueva información.

2. **Verosimilitud**: son las probabilidades de obtener un conjunto de datos, dados ciertos valores de los parámetros iniciales.

3. **Distribución Posteriori**: combina la distribución priori y la verosimilitud para proporcionar una actualización sobre los parámetros después de observar los datos obtenidos.

La ecuación de regresión en modelos lineales y bayesianos

La ecuación de regresión bayesiana y la estándar se diferencian en los términos presentados anteriormente. En esta sección se presentará el modelo lineal y luego el bayesiano, para establecer comparaciones entre ellos con dos variables independientes.

La regresión lineal múltiple con dos variables X_1 y X_2, se puede expresar con la siguiente ecuación:

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Dónde: Y es la variable dependiente predicha (respuesta); β_0 es el intercepto β_1 y β_2 son los coeficientes de las variables independientes x_1 y x_2 respectivamente, y ϵ es el término de error, que se asume que sigue una distribución normal con media 0 y varianza σ^2: $\epsilon \sim N(0,\sigma^2)$

La estimación de los parámetros β_0, β_1, β_2, se realiza minimizando la suma de los cuadrados de los errores (OLS - *Ordinary Least Squares*):

$$\hat{\beta} = argim \sum_{i=1}^{n} \left(Y_i - \widehat{Y_i}\right)^2$$

La expresión *argmin* (argumento del mínimo), indica el valor del argumento que minimiza la función. En la regresión lineal se utiliza el método de los mínimos cuadrados ordinarios (OLS) para encontrar los valores de los parámetros que minimizan la suma de los cuadrados de los errores (residuos). En otras palabras, la función que queremos minimizar es la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos, siendo estos últimos los obtenidos por el modelo de regresión.

En la regresión bayesiana, en lugar de estimar un único valor para los parámetros, se estiman distribuciones a posteriori para los mismos. Se utiliza el teorema de Bayes para actualizar nuestras creencias sobre los parámetros del modelo, dado los datos observados, y esto se realiza en varios pasos:

Paso 1: Definir el Modelo y la Verosimilitud. Aquí el modelo lineal no cambia:

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$\sigma^2: \epsilon \sim N(0,\sigma^2)$$

Paso 2: Especificar las distribuciones a priori. Se asignan las distribuciones a priori para los parámetros β_0, β_1, β_2 y σ^2:

$\beta_0 \sim N(\mu_{\beta 0}, \sigma^2{}_{\beta 0})$

$\beta_1 \sim N(\mu_{\beta 1}, \sigma^2{}_{\beta 1})$

$\beta_2 \sim N(\mu_{\beta 2}, \sigma^2{}_{\beta 2})$

$\sigma^2 \sim$ Gamma Inversa (α, β)

Paso 3: Calcular la Distribución Posteriori. En este paso se utiliza el teorema de Bayes para combinar la verosimilitud y las distribuciones a priori, y así obtener la distribución posteriori de los parámetros:

$p\ (\beta_0, \beta_1, \beta_2, \sigma^2 \mid \text{datos}) \propto p(\text{datos} \mid \beta_0, \beta_1, \beta_2, \sigma^2) * p(\beta_0) * p(\beta_1) * p(\beta_2) * p(\sigma^2)$

Donde:

$p(\text{datos} \mid \beta_0, \beta_1, \beta_2, \sigma^2)$ es la verosimilitud.

$p(\beta_0)$, $p(\beta_1)$, $p(\beta_2)$, $p(\sigma_2)$ son las distribuciones a priori.

Nótese que la función de probabilidad asignada a la varianza, es la gama inversa, en contraste con la normal estándar. La función gama inversa es una distribución de probabilidad continua que se utiliza especialmente en la inferencia bayesiana.

Por definición, si una variable aleatoria X tiene una distribución gama inversa con parámetros α (forma) y β (escala), se denota como $X \sim$ Inv-Gamma (α, β), si su función de densidad de probabilidad es:

$$\mathrm{f}(\mathrm{x} \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-(\alpha-1)} e^{-\beta/x}$$

Para x>0, donde $\Gamma\ (\alpha)$ es la función gamma.

La distribución gama inversa, tiene las siguientes medidas de tendencia central y dispersión:

Esperanza (media): $E\ [X]= \beta/\alpha-1$, para α>1

Varianza: Var $(X)=\beta^2 / (\alpha - 1)^2\ (\alpha -2)$, para α>2

En el contexto de la regresión bayesiana, uno de los principales objetivos es estimar los parámetros del modelo: los coeficientes β y la varianza del error σ^2. Como se mostró, la función gama inversa se usa comúnmente como una distribución a priori para la varianza del error en la regresión bayesiana por varias razones: a) Conjugación: la distribución gama inversa es la distribución a priori conjugada para la varianza en modelos lineales gaussianos. Esto significa que, si la varianza del error σ^2 tiene una distribución gama inversa a priori, la distribución a posteriori de σ^2 después de observar los datos también será una gama inversa. Esta propiedad de conjugación simplifica los cálculos analíticos y permite obtener resultados de forma más eficiente; b) Flexibilidad: la distribución gama inversa es flexible y puede modelar una amplia gama de distribuciones para σ^2, dependiendo de los parámetros α y β, lo que permite incorporar diferentes niveles de conocimiento previo sobre σ^2; c) simplicidad matemática: la forma de la distribución gama inversa hace que sea conveniente para el muestreo y la implementación en algoritmos bayesianos.

La propiedad de conjugación en estadística bayesiana se refiere a la situación en la que la forma de la distribución a priori y la forma de la distribución a posteriori son de la misma familia. Esta distribución combina nuestra creencia previa (a priori) con la información de los datos observados. Por lo tanto, si una distribución a priori y una distribución a posteriori son "conjugadas", significa que, si la distribución a priori pertenece a una cierta familia de distribuciones, la distribución a posteriori también pertenecerá a esa misma familia.

Volvamos por un momento al paso 3 que requiere calcular la distribución a posteriori. Habíamos dicho que en este paso se utiliza el teorema de Bayes para combinar la verosimilitud y las distribuciones a priori, y así obtener la distribución posteriori de los parámetros:

$p\ (\beta_0, \beta_1, \beta_2, \sigma^2 \mid \text{datos}) \propto p(\text{datos} \mid \beta_0, \beta_1, \beta_2, \sigma^2) * p(\beta_0) * p(\beta_1) * p(\beta_2) * p(\sigma^2)$

Esto significa que la distribución posteriori de los parámetros es proporcional al producto de la verosimilitud de los datos, dados los parámetros y las distribuciones a priori de los parámetros. Aquí es donde se aplica el Teorema de Bayes y la Proporcionalidad, el cual establece que:

$$p(\theta \mid \text{datos}) = \frac{p(datos|\theta) * p(\theta)}{P(datos)}$$

Donde:

$p(\theta \mid \text{datos})$ es la distribución posteriori de los parámetros θ dados los datos;

$p(\text{datos} \mid \theta)$ es la verosimilitud de los datos dados los parámetros θ;

$p(\theta)$ es la distribución a priori de los parámetros θ;

$p(\text{datos})$ es la verosimilitud marginal o evidencia, que es una constante de normalización;

En el contexto de la regresión bayesiana con los parámetros β_0, β_1, β_2, σ^2, el teorema se representa por la siguiente ecuación:

$$p(\beta_0, \beta_1, \beta_2, \sigma^2|datos) = \frac{p(datos|\beta_0, \beta_1, \beta_2, \sigma^2) * p(\beta_0) * p(\beta_1) * p(\beta_2) * p(\sigma^2)}{p(datos)}$$

El término $p(\text{datos})$, es conocido como la constante de proporcionalidad o la evidencia; es el mismo para todos los valores posibles de los parámetros y garantiza que la distribución posteriori se normalice adecuadamente para que la integral de la distribución posteriori sobre todo el espacio de parámetros sea igual a 1.

$p\ (\beta_0, \beta_1, \beta_2, \sigma^2 \mid \text{datos}) \propto p(\text{datos} \mid \beta_0, \beta_1, \beta_2, \sigma^2) * p(\beta_0) * p(\beta_1) * p(\beta_2) * p(\sigma^2)$

Matemáticamente, α en la expresión de proporcionalidad se refiere a la constante de normalización que asegura que la distribución posteriori se integre a 1. La regresión bayesiana, se enfoca en la relación proporcional para entender cómo se combinan la

verosimilitud y las distribuciones a priori, para formar la distribución posteriori de los parámetros.

Ejemplo: Calificación Final Basada en Horas de Estudio y Asistencia

Un grupo de investigadores desea predecir la calificación final de una muestra de alumnos, en una prueba de geometría. Selecciona como variables predictoras de la calificación final, las horas dedicadas al estudio y la asistencia a clases. En el análisis descriptivo detectan que las dos variables independientes (predictoras) se hallan altamente correlacionadas, dado que buena parte del tiempo de estudio se realizó durante las horas de clase. Por otro lado, el tamaño muestral es pequeño. Dados estos inconvenientes deciden aplicar una regresión lineal bayesiana, antes que una estándar. El ejemplo propuesto, se realizó con ayuda del software estadístico JASP. Los resultados se muestran a continuación.

Datos:

Variable dependiente (Y): Calificación final

Variables independientes (X_1, X_2): X_1: Horas de estudio; X_2: Asistencia.

En la siguiente imagen se muestra la matriz de datos cargada en el software JASP

	Horas de estudio	Asistencia	Calif Final
1	10	80	90
2	8	70	85
3	15	90	95
4	5	60	70
5	12	85	65

Una vez cargados los datos, procedemos a realizar el análisis de regresión bayesiano. La primera tabla que ofrece el software es la comparación de modelos.

Comparación de Modelos - Calif Final

Modelos	P(M)	P(M|datos)	FB_M	FB_{10}	R^2
Modelo nulo	0.333	0.411	1.398	1.000	0.000
Horas de estudio + Asistencia	0.333	0.291	0.822	0.708	0.230
Horas de estudio	0.167	0.172	1.038	0.835	0.135
Asistencia	0.167	0.125	0.717	0.610	0.073

¿Qué es el Modelo Nulo? El modelo nulo es el modelo más simple posible en un análisis de regresión. Este no incluye ninguna variable independiente (predictoras: horas de estudio y/o asistencia). En otras palabras, asume que la calificación final no depende de ninguna variable explicativa y predice la misma calificación para todos los estudiantes, basada únicamente en el promedio de las calificaciones observadas.

El modelo nulo se expresa matemáticamente de la siguiente manera:

$Y=\beta_0+\epsilon$

donde:

β_0: es el promedio de la calificación final para todos los estudiantes.

ϵ: es el término de error (las diferencias entre las calificaciones individuales y el promedio).

Entonces, el modelo nulo predice la misma calificación para todos los estudiantes, sin importar cuántas horas hayan estudiado o cuánto hayan asistido a clases. Su predicción es simplemente el promedio de las calificaciones de los estudiantes en el conjunto de datos. El modelo nulo sirve como punto de referencia para comparar otros modelos. Al no incluir ninguna variable independiente, proporciona un valor base de R^2, el Factor Bayes (FB), y otras métricas para evaluar cuánto mejoran los modelos cuando se incluyen los predictores. En el ejemplo, la comparación que se realiza en el renglón siguiente al modelo nulo, incluye las variables independientes predictoras (horas de estudio y asistencia), y se aprecia una mejora en relación con el modelo nulo, ya que el coeficiente de determinación es mayor R^2= 0.225, entonces se concluye que las variables predictoras aportan valor para explicar la variabilidad en las calificaciones finales. Los renglones siguientes, representan los modelos con una sola variable predictora, horas de estudio o asistencia, y se observa que R^2 aumenta, aunque no en la misma magnitud que el modelo que combina los dos predictores. R^2 del modelo nulo es cero porque no explica ninguna variabilidad en la calificación final, ya que simplemente predice el promedio. Comparar el R^2 de los modelos con variables predictoras respecto al modelo nulo muestra cuánto de la variabilidad en las calificaciones es explicado por las horas de estudio y la asistencia.

En la columna Factor Bayes (FB), se tiene que, si el valor FB_{10} de un modelo (con horas de estudio, asistencia, o ambos) es mayor que 1 en relación con el modelo nulo, significa que hay evidencia a favor de incluir las variables predictoras para explicar la calificación final. Un FB_{10} cercano a 1 indica que el modelo con predictores no mejora sustancialmente respecto al modelo nulo. En el ejemplo, se tiene que ningún valor del Factor Bayes es mayor que 1. Por lo tanto, cuando los coeficientes Factor Bayes (FB) son menores que 1 al incluir las variables predictoras (en este caso, horas de estudio y asistencia) en el modelo, la interpretación es que el modelo que incluye esas variables no mejora el ajuste con respecto al modelo nulo. De hecho, sugiere que el modelo nulo podría ser preferible.

Recordemos que el Factor Bayes (FB) BF_{10} es la relación de verosimilitud entre dos modelos, usualmente el modelo de interés (con variables predictoras) y el modelo nulo (sin variables predictoras).

Si FB_{10}>1 indica evidencia a favor del modelo con predictores frente al modelo nulo.

Si FB_{10}<1 indica evidencia a favor del modelo nulo, lo que sugiere que añadir las variables predictoras no mejora el ajuste del modelo.

En este ejemplo, al incluir las variables independientes: horas de estudio y asistencia, los coeficientes FB_{10} son menores que 1, esto significa que el modelo nulo (que solo predice el promedio de las calificaciones finales) es preferible al modelo que incluye una

o ambas variables predictoras. Esto puede deberse a varias razones. Las variables predictoras no explican bien la variabilidad en la calificación final: horas de estudio y asistencia podrían no tener una relación significativa con la calificación final en los datos, o esta relación es muy débil. En términos estadísticos, las variables predictoras no aportan información sustancial adicional sobre la calificación final comparado con el simple promedio que asume el modelo nulo. Otra posibilidad es que el modelo con predictores podría estar sobre ajustado. Si el modelo incluye demasiados parámetros o variables irrelevantes, podría ajustarse demasiado a los datos de entrenamiento, lo que da como resultado un peor desempeño cuando se compara con el modelo nulo, que es más simple.

En un análisis bayesiano, el Factor Bayes también refleja la incertidumbre sobre si los coeficientes de las variables predictoras son realmente diferentes de cero (es decir, si tienen un impacto en la variable dependiente). Si esta incertidumbre es alta y la evidencia para un impacto positivo de las variables es baja, el FB_{10} será menor que 1.

En los siguientes renglones de la tabla, se listan los diferentes modelos ajustados a los datos.

Modelo Completo: $Y=\beta_0+\beta_1X_1+\beta_2X_2$ (ambas variables, horas de estudio y asistencia).

Modelo 1: $Y=\beta_0+\beta_1X_1$ (solo horas de estudio).

Modelo 2: $Y=\beta_0+\beta_2X_2$ (solo asistencia).

La comparación entre los modelos permite evaluar el impacto de incluir o excluir variables.

En la tabla de Comparación de Modelos de JASP, las columnas P(M) y P(M|datos) son elementos clave en la interpretación bayesiana. Vamos a desglosar lo que significa cada una:

Probabilidad a Priori del Modelo P(M): representa la probabilidad a priori asignada a cada modelo antes de ver los datos. Esta es la probabilidad que se le otorga al modelo basado solo en un conocimiento previo, sin observar los datos. En muchos análisis bayesianos, si no se tienen razones específicas para preferir un modelo sobre otro, el software JASP asume probabilidades a priori iguales para todos los modelos. Por lo tanto, en muchos casos, los valores de P(M) serán iguales para todos los modelos, distribuyendo la probabilidad de manera equitativa entre ellos. Si se tiene información adicional sobre un modelo en particular (por ejemplo, resultados de estudios previos que sugieren que cierto modelo es más plausible), se ajustan las probabilidades a priori de manera diferente.

Probabilidad a Posteriori del Modelo P(M|datos): es la probabilidad a posteriori de cada modelo, es decir, la probabilidad de que ese modelo sea el mejor, después de haber observado los datos. Esta columna es fundamental en el análisis bayesiano, ya que refleja cómo cambian nuestras creencias acerca de cada modelo una vez que hemos tomado en cuenta la evidencia de los datos. Está calculada utilizando el Teorema de Bayes, que actualiza la probabilidad a priori $P(M)$ de cada modelo con la información de los datos.

La fórmula del teorema de Bayes aplicada a los modelos es:

$$P(M|datos) = \frac{P(datos\ |M) * P(M)}{\sum_j P(datos\ |M_j) * P(M_j)}$$

Donde:

$P(M|\text{datos})$ es la probabilidad a posteriori del modelo dado los datos.

$P(\text{datos}|M)$ es la verosimilitud del modelo (qué tan bien explica los datos observados).

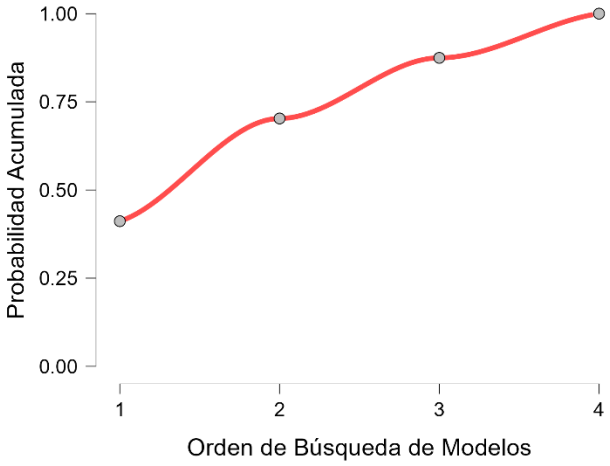
$P(M)$ es la probabilidad a priori del modelo.

El denominador es la suma de las verosimilitudes ponderadas de todos los modelos, lo que normaliza las probabilidades. Si uno de los modelos tiene un valor de P(M|datos) mayor que los otros, significa que, dados los datos, este modelo es mucho más probable que los demás. Un valor de P(M|datos) cercano a 1 indica que el modelo es fuertemente apoyado por los datos.

En este ejemplo, P(M), esto es, las probabilidades a priori, se reparten de manera proporcional entre el modelo nulo y el modelo completo. La suma de las probabilidades de los modelos con un solo predictor es igual al modelo completo dividido dos. Si no hay un conocimiento previo que favorezca algún modelo, se asignan probabilidades a priori iguales. En este caso 0,333+0,333+0,167+0,167= 1 (sumando las probabilidades asignadas a todos los modelos de la tabla).

Después de analizar los datos, la columna P(M|datos) mostrará cómo las probabilidades han cambiado. Aquí, la probabilidad a posteriori depende de la precisión con que cada modelo se ajusta los datos. En el ejemplo, para el modelo nulo (M0), P(M|datos) = 0.411. Esto significa que, después de analizados los datos, hay un 41,1% de probabilidad de que el modelo nulo sea el mejor. Para el modelo completo, P(M|datos) = 0.291, es decir, el modelo tiene una probabilidad de un 29,1% de ser el mejor modelo. La interpretación es similar para los modelos con un predictor, sean las horas de estudio o la asistencia.

Gráficas de Probabilidades del Modelo



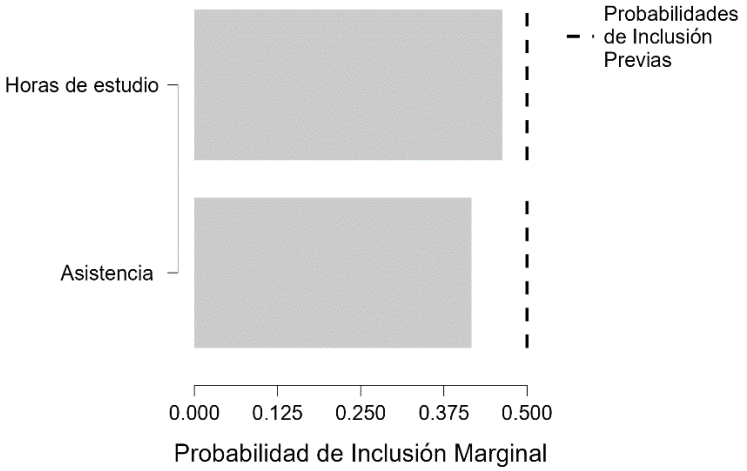
En el gráfico denominado probabilidades del modelo que ofrece JASP, el eje X es el orden de búsqueda de los modelos, y el eje Y es la probabilidad acumulada. Proporciona una representación visual de cómo se distribuyen las probabilidades a posteriori $P(M|datos)$ entre los distintos modelos en el análisis bayesiano. Para interpretar la gráfica, vamos a partir del eje X: orden de búsqueda de los modelos, que representa el

orden en que se consideran los modelos en la búsqueda. Generalmente, en los análisis bayesianos, se evalúan varios modelos, desde los más simples (el modelo nulo) hasta los más complejos (el modelo completo). Cada punto en el eje X corresponde a un modelo específico que ha sido evaluado. En el eje Y se muestra la probabilidad acumulada, que es la suma de las probabilidades a posteriori $P(M|datos)$ de todos los modelos evaluados hasta ese punto. Esto significa que, al moverse de izquierda a derecha en el gráfico, se suman las probabilidades de los modelos en ese orden, hasta alcanzar una probabilidad acumulada de 1 (100% de probabilidad).

La curva en el gráfico muestra cómo las probabilidades a posteriori se distribuyen entre los modelos a medida que estos se consideran en orden. Modelos con alta probabilidad a posteriori tendrán un mayor impacto en el crecimiento de la probabilidad acumulada en el eje Y. Si un modelo tiene una probabilidad $P(M|datos)$ significativamente alta, la curva tendrá un salto más pronunciado cuando ese modelo es evaluado. Al contrario, modelos con baja probabilidad a posteriori tendrán un menor impacto en la curva, y la acumulación será más gradual cuando se incluyan esos modelos. Esta última situación es la que se presenta en el ejemplo.

Por lo tanto, los modelos que tienen un impacto más significativo en la curva son aquellos con las probabilidades a posteriori más altas. Si la curva presenta un gran aumento en la probabilidad acumulada en un punto determinado, significa que el modelo correspondiente en ese punto tiene una fuerte evidencia de ser el mejor modelo según los datos. Si la curva se estabiliza rápidamente (es decir, llega a una probabilidad acumulada cercana a 1 después de evaluar pocos modelos), esto indica que uno o pocos modelos dominan el conjunto de modelos en términos de probabilidad.

Gráfica de Probabilidades de Inclusión



En el gráfico de Probabilidades de inclusión que se ofrece un análisis bayesiano de regresión, donde se muestra la probabilidad de que cada variable predictora (independiente) esté incluida en el mejor modelo. Este gráfico es útil para evaluar la relevancia de cada predictor (horas de estudio y asistencia) en la predicción de la variable dependiente (calificación final).

El gráfico de Probabilidades de inclusión muestra la fuerza de la evidencia para que una variable predictora deba estar incluida en el modelo. Cada variable tiene una barra que muestra la probabilidad de que esté presente en el modelo que mejor explica los datos.

Una probabilidad cercana a 1 (o al 100%) significa que es muy probable que esa variable esté incluida en el mejor modelo, lo que sugiere que esa variable tiene un impacto relevante en la predicción de la variable dependiente. Al contrario, probabilidades cercanas a 0 significa que es poco probable que esa variable esté incluida en el mejor modelo, lo que indica que su contribución a la predicción es débil o insignificante.

Para horas de estudio, la probabilidad de inclusión es de p=0,463 (46,3%), lo que indica que esta variable no tiene un fuerte impacto en la predicción de la calificación final. Para asistencia, la barra correspondiente también indica una baja probabilidad de inclusión p= 0,417 (41,7%), lo que sugiere que esta variable no es particularmente útil para predecir las calificaciones finales.

El gráfico de probabilidades de inclusión ayuda a identificar las variables más relevantes para incluir en el modelo final: variables con alta probabilidad de inclusión deben ser consideradas esenciales en el modelo porque aportan información valiosa. Variables con baja probabilidad de inclusión podrían ser descartadas o incluidas con menor prioridad, ya que no aportan significativamente a la predicción.

Tabla de Resumen Posterior

Resúmenes Posteriores de los Coeficientes

								Intervalo con el 95% de Credibilidad	
Coeficiente	**P(incl)**	**P(excl)**	**P(incl|datos)**	**P(excl|datos)**	**$FB_{inclusión}$**	**Media**	**DT**	**Inferior**	**Superior**
Intercept	1.000	0.000	1.000	0.000	1.000	80.467	3.279	72.965	87.819
Horas de estudio	0.500	0.500	0.463	0.537	0.863	1.223	2.251	-1.140	7.519
Asistencia	0.500	0.500	0.417	0.583	0.714	-0.257	0.734	-2.078	0.883

La tabla de Resúmenes Posteriores de los Coeficientes en el software JASP brinda información clave sobre la importancia y el impacto de cada variable predictora (independiente) en la predicción de la variable dependiente. Según el orden en que aparecen las columnas estas indican:

P(incl): Es la probabilidad a priori de que una variable sea incluida en el modelo, es decir, la probabilidad de que esa variable tenga un impacto significativo en la predicción antes de observar los datos. Este valor refleja el conocimiento que se tiene antes de ver los datos y, en muchos casos, puede estar fijado en 0.5 si no se tiene información previa específica (lo que significa que se asume que es igualmente probable que la variable sea incluida o no).

P(excl): Es la probabilidad a priori de que una variable sea excluida del modelo. Matemáticamente es complementaria a $P(incl)$, es decir, $P(excl)$=1−$P(incl)$. Si $P(incl)$=0.5, entonces $P(excl)$ también sería 0.5, lo que significa que, antes de observar los datos, no se tiene preferencia entre incluir o excluir la variable.

P(incl|datos): Es la probabilidad a posteriori de que una variable esté incluida en el modelo, dado los datos observados. Esta probabilidad se calcula utilizando el Teorema de Bayes y refleja cuánta evidencia aportan los datos para apoyar la inclusión de la variable en el modelo. Un valor alto de P(incl|datos) (por ejemplo, cercano a 1) indica

que los datos sugieren fuertemente que esta variable es relevante para predecir la variable dependiente (en este caso, la calificación final). Un valor bajo (cercano a 0) sugiere que los datos no respaldan la inclusión de esa variable en el modelo.

P(excl|datos): Es la probabilidad a posteriori de que una variable sea excluida del modelo, dado los datos observados. Es complementaria a $P(incl|datos)$, es decir, $P(excl|datos)$=1−$P(incl|datos)$. Un valor alto de P(excl|datos) sugiere que los datos respaldan la exclusión de esa variable, indicando que probablemente no es necesaria para la predicción de la variable dependiente.

En el ejemplo, las probabilidades de excluir ambas variables predictoras son más altas que incluirlas. Lo cual refleja que no se ha ganado nada respecto del modelo nulo.

FB inclusión (Factor Bayesiano de Inclusión): Es una medida que cuantifica la evidencia de los datos a favor de la inclusión de la variable en el modelo, comparada con su exclusión. Como regla general se considera que: si FB inclusión>1, los datos proporcionan evidencia a favor de la inclusión de la variable en el modelo. Si FB inclusión< 1, los datos sugieren que es más probable que la variable no sea necesaria en el modelo. El Factor Bayesiano es una de las principales herramientas en el análisis bayesiano para evaluar la fuerza de la evidencia en comparación con un modelo nulo.

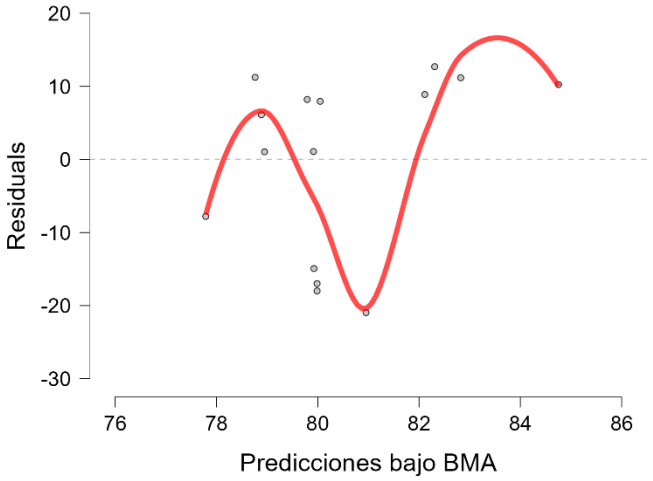
En el ejemplo los valores de FB inclusión de las variables independientes, están próximos a 0,5 indicando que los datos no aportan evidencia clara ni para la inclusión ni para la exclusión.

Media: En esta columna, se presenta el valor promedio del coeficiente posterior de cada predictor (variable independiente), dado los datos observados. Este valor representa la magnitud y la dirección del efecto que tiene la variable sobre la variable dependiente. Un valor positivo sugiere que un aumento en esa variable predictora está asociado con un aumento en la variable dependiente (calificación final). Un valor negativo indica que un aumento en esa variable predictora está asociado con una disminución en la variable dependiente.

En el ejemplo la media para la variable asistencia es negativa, esto significa que, dado los datos, hay evidencia de que mayor asistencia podría estar asociada con calificaciones finales más bajas, lo cual puede ser un hallazgo contraintuitivo, pero como el modelo completo no supera al modelo nulo, no se interpretaría esta tendencia en los datos.

DT (Desviación Típica): Es la desviación estándar del coeficiente posterior de cada predictor. Indica la incertidumbre asociada con la estimación de la media del coeficiente posterior. Un valor alto de DT sugiere que hay mayor incertidumbre sobre el valor verdadero del coeficiente (es decir, el rango en el que podría estar el valor verdadero del coeficiente es más amplio). Un valor bajo de DT indica que hay menos incertidumbre, y el valor del coeficiente está mejor definido. La combinación de la media y la desviación típica da una idea de cuánto afecta cada variable a la predicción de la calificación final y con cuánta certeza puedes confiar en esa estimación.

Gráfica de Errores vs. Ajustado



El gráfico Errores vs. Ajustado que ofrece JASP en un análisis de regresión es una herramienta diagnóstica para evaluar la calidad del ajuste del modelo. En este gráfico se representa la relación entre los valores ajustados (las predicciones del modelo) y los errores residuales (las diferencias entre los valores observados y los valores predichos por el modelo). El eje x representa los valores ajustados o predicciones del modelo. Estos son los valores que el modelo estima para la variable dependiente (en este caso, la calificación final) en función de las variables independientes (horas de estudio y asistencia). El eje y representa los errores residuales, es decir, las diferencias entre los valores reales observados de la variable dependiente y los valores predichos por el modelo:

Error residual= Yobservado$-Y$ajustado

Patrones esperados en el gráfico:

Distribución aleatoria de los errores: Un buen ajuste de modelo debería mostrar una dispersión aleatoria de los errores alrededor de cero en el gráfico. Esto significa que los errores no deben mostrar patrones sistemáticos, sino estar distribuidos sin tendencia aparente en función de los valores ajustados. Esto sugiere que el modelo está capturando adecuadamente la relación entre las variables independientes y la dependiente.

Sin tendencia o forma particular: Si los puntos en el gráfico están distribuidos uniformemente alrededor de cero, sin formar patrones, esto indica que el modelo ajusta bien los datos y que los errores son independientes.

Patrones que muestran sesgos:

Patrón en forma de U o U invertida: Si los puntos forman una curva en forma de U o una parábola, esto podría indicar que el modelo no está capturando correctamente algún patrón no lineal en la relación entre las variables independientes y la dependiente. En este caso, podría ser necesario explorar una transformación de las variables o considerar un modelo que pueda captar mejor la no linealidad.

Distribución no homogénea de los errores (heterocedasticidad): Si los puntos en el gráfico muestran una dispersión que aumenta o disminuye sistemáticamente a lo largo de los valores ajustados (por ejemplo, si los puntos están más dispersos para valores altos o bajos de los ajustados), esto indica heterocedasticidad. La heterocedasticidad

sugiere que el modelo tiene problemas con la consistencia de los errores a lo largo del rango de los valores ajustados, lo que puede afectar la fiabilidad de los resultados del modelo.

Errores sistemáticos: si el gráfico muestra algún patrón o tendencia clara (por ejemplo, una curva), esto indica que el modelo no está capturando correctamente la relación entre las variables independientes y la dependiente. Es una señal de que puede haber un sesgo estructural en el modelo o que faltan variables importantes.

En el ejemplo que hemos desarrollado, no se observan sesgos en los datos, por lo tanto, se concluye que tanto las horas dedicadas al estudio y la asistencia no aportan de manera significativa a comprender las calificaciones finales. En otras palabras, el mejor predictor de las mismas es la media de calificaciones del grupo.

Reconstruyendo el ejemplo

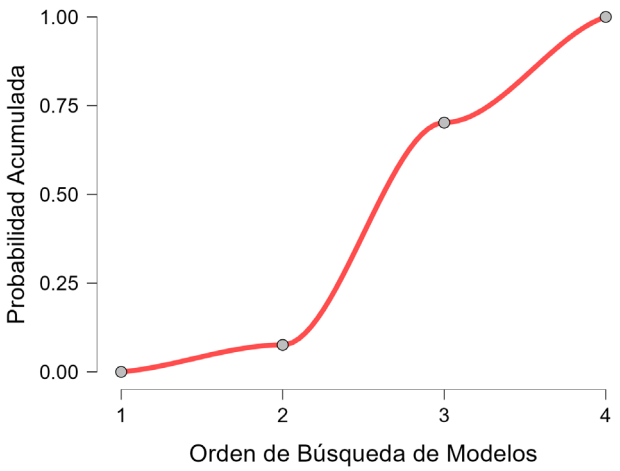
Los investigadores resultaron decepcionados al comprobar que no pudieron predecir la calificación final de una muestra de alumnos en una prueba de geometría. Supusieron adecuadamente que la calificación final, dependía de las horas dedicadas al estudio y la asistencia a clases. Pero en ese supuesto dieron por sentado que las dos variables independientes (predictoras) se hallan altamente correlacionadas, entendiendo que buena parte del tiempo de estudio se realizó durante las horas de clase. Una revisión más exhaustiva de los datos demostró que una proporción importante de las horas de estudio tenía lugar en la casa a través de las tareas que los profesores les daban a los escolares. Asimismo, cambiaron el modo de calificación de la prueba por una nota en vez de porcentaje de aprobación. Con estas modificaciones, se propusieron revisar el modelo de regresión, pero dado que no pudieron aumentar el tamaño muestral, aplicaron nuevamente una regresión lineal bayesiana. Los resultados de este nuevo análisis se muestran a continuación.

Comparación de Modelos - Calif Fin

Modelos	P(M)	P(M|datos)	FB_M	FB_{10}	R^2
Hs de estudio	0.167	0.626	8.375	1.000	0.846
Hs de estudio + Asiste	0.333	0.298	0.848	0.238	0.856
Asiste	0.167	0.076	0.411	0.121	0.784
Modelo nulo	0.333	6.283×10^{-5}	1.257×10^{-4}	5.017×10^{-5}	0.000

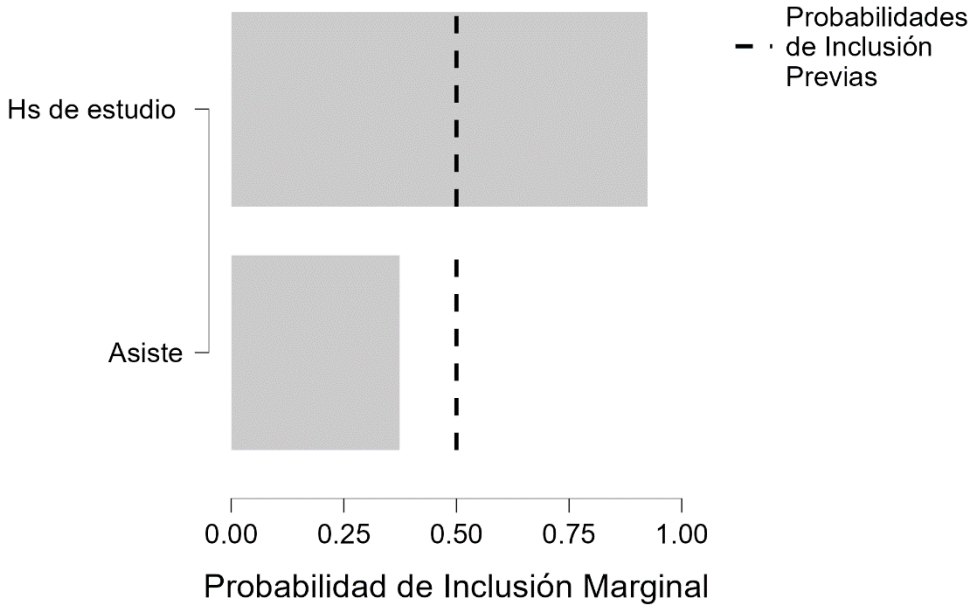
En la tabla que compara los diferentes modelos, se puede apreciar ahora que el Factor de Bayes FB_{10} que incluye horas de estudio, alcanza la unidad y se diferencia de los otros modelos, especialmente del modelo nulo. La proporción de varianza explicada por es R^2= 0,846 (84,6%), que solo es superada ligeramente por el modelo completo (horas de estudio + asistencia). Por lo tanto, en este caso tendríamos un modelo preferible en comparación con el modelo nulo, el incluye la variable horas de estudio. Asimismo, se podría descartar que la variable Asistencia tenga un impacto significativo sobre el rendimiento en la evaluación. Lo dicho se corrobora observando los cambios entre las probabilidades a priori y las probabilidades a posteriori (columnas P(M) y P(M|datos)). La conclusión que se extrae de la columna P(M|datos) es que el modelo que incluye horas de estudio tiene una probabilidad asociada de 62,6% sea el que explica mejor el comportamiento de la variable dependiente (calificaciones).

Gráficas de Probabilidades del Modelo



Recordemos que esta gráfica representa visualmente cómo se distribuyen las probabilidades a posteriori $P(M|datos)$ entre los distintos modelos en el análisis bayesiano. Se observa que al menos un modelo produce un cambio abrupto en la pendiente de la curva, y por los datos expuestos en la tabla anterior, se reconoce que el cambio de tendencia se logra cuando se incluye el modelo con la variable horas de estudio. Sería este entonces, el modelo con el impacto más importante.

Gráfica de Probabilidades de Inclusión



Se muestra aquí la probabilidad de que cada variable predictora (independiente) esté incluida en el mejor modelo. Como ya se explicó, el gráfico es útil para evaluar la relevancia de cada predictor (horas de estudio y asistencia) en la predicción de la

variable dependiente (calificación final). Claramente surge de esta visualización que horas de estudio es la variable con mayor probabilidad de ser incluida: p= 0,924.

Resumen Posterior

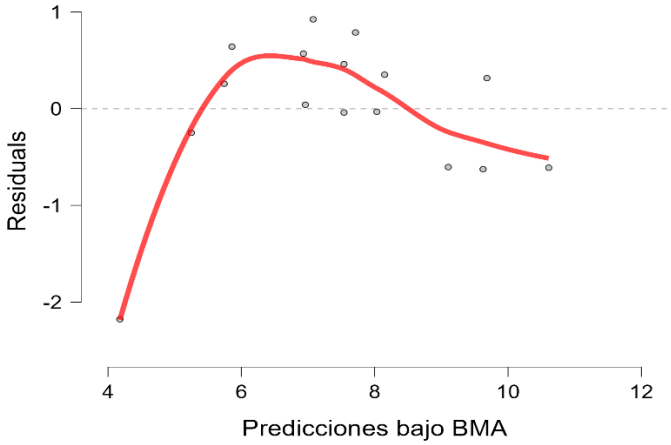
Resúmenes Posteriores de los Coeficientes

Coeficiente	P(incl)	P(excl)	P(incl|datos)	P(excl|datos)	$FB_{inclusión}$	Media	DT	Intervalo con el 95% de Credibilidad	
								Inferior	Superior
Intercept	1.000	0.000	1.000	0.000	1.000	7.500	0.204	7.076	7.930
Hs de estudio	0.500	0.500	0.924	0.076	12.163	0.463	0.180	0.000	0.657
Asiste	0.500	0.500	0.374	0.626	0.597	0.029	0.058	-0.023	0.188

Con los datos de esta tabla se corrobora que el modelo que mejor explica el resultado del examen (calificaciones como variable dependiente) son las horas de estudio, dados los valores de P(incl|datos)= 0,924 y FB inclusión= 12,163 > 1.

En la última etapa se analiza el grafico de residuos ajustado para comprobar si no existen patrones evidentes en la distribución de los errores,

Gráfica de Errores vs. Ajustado



Se observa que para casi todos los casos, la distribución de los residuos se encuentra cerca de cero, excepto para un caso que representa un dato distante (*outlier*). En este caso, la sugerencia es estudiar dicho caso en particular para verificar si se trata de un caso al que se deba prestar atención, o un error de codificación.

Síntesis

Se han propuesto dos ejemplos para comprender los fundamentos de la regresión bayesiana como enfoque estadístico útil para modelar y analizar la relación entre variables dependientes e independientes usando el teorema de Bayes. La utilidad de este modelo de regresión en comparación con la regresión lineal, radica en que la

regresión bayesiana tiene la ventaja de incorporar nueva información sobre los parámetros del modelo a medida que obtenemos nuevos datos. Si se tiene información relevante de los parámetros, previo a una investigación, es posible utilizar esa información a priori para determinar las propiedades explicativas del modelo a medida que se suma nueva información. Por otra parte, es menos restrictiva en que la regresión lineal en cuanto a los supuestos subyacentes. Asimismo, cuando se trabaja con tamaños de muestra pequeños, suele ser más útil dado que se distorsionan menos los valores de las estimaciones. Por último, para la elaboración de este trabajo se utilizó el software JASP, los motivos de su elección se expresan en la filosofía de los autores del software que se resumen de la siguiente manera: JASP es un proyecto de código abierto apoyado por la Universidad de Ámsterdam, tiene una interfaz intuitiva que fue diseñada pensando en el usuario y ofrece procedimientos de análisis estándar tanto en su forma clásica como bayesiana. Quienes estén interesados en trabajar con JASP pueden consultar su página oficial:

JASP https://jasp-stats.org/