Práctica 4

Regresión lineal

El objetivo principal de esta unidad es la aplicación de modelos de regresión lineales desde una perspectiva bayesiana, considerando a los parámetros del modelo como cantidades aleatorias que se corresponden con una distribución de probabilidad *a priori*. A diferencia del enfoque frecuentista o máximo verosímil, el resultado de la inferencia bayesiana es una distribución de probabilidad *a posteriori*, la cual se utiliza como fuente de todas las conclusiones. Además, se emplean técnicas propias de la estadística bayesiana para evaluar la adecuación y comparar los modelos utilizados.

1. Mi primer regresión bayesiana

El conjunto de datos `sales` contiene los montos semanales de inversión en publicidad y de ingresos de una determinada compañía. Considere el siguiente modelo de regresión lineal simple:

$$\begin{aligned} \text{ventas}_i &\sim \text{Normal}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 \text{publicidad}_i \end{aligned}$$

i. Ajuste el modelo utilizando `{Rstan}` y distribuciones uniformes como *priors*.
ii. Construya un gráfico que muestre las ventas en función de la inversión en publicidad y superponga la recta de regresión estimada.

2. Mejorando mi regresión bayesiana

Considere la siguiente versión del modelo del ejercicio anterior que propone distribuciones *a priori* para los parámetros del modelo:

$$\begin{aligned} \text{ventas}_i &\sim \text{Normal}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 \text{publicidad}_i \\ \beta_0 &\sim \text{Normal}(\overline{\text{ventas}}, 10^2) \\ \beta_1 &\sim \text{Normal}(0, 0.5^2) \\ \sigma &\sim \text{Normal}^+(5) \end{aligned}$$

i. Ajuste el modelo utilizando `{RStan}` y los *priors* sugeridos.
ii. Construya un gráfico que muestre las ventas en función de la inversión en publicidad, superponga la recta de regresión estimada, y el intervalo de credibilidad del 95% para la recta de regresión.

3. Regresiones frecuentistas y bayesianas

Utilice datos simulados para comparar la estimación por mínimos cuadrados con la estimación Bayesiana en modelos de regresión.

i. Simule 100 observaciones del modelo $Y = 2 + 3X + \varepsilon$ donde los valores del predictor X se obtienen de una distribución Uniforme$(0, 20)$ y los errores son obtenidos de manera independiente de una distribución Normal$(0, 5^2)$.
ii. Ajuste el modelo de regresión utilizando `lm()` y un modelo bayesiano mediante `{RStan}` utilizando *priors* uniformes.

iii. Verifique que ambos métodos arrojan resultados similares.
iv. Represente gráficamente los datos y las dos rectas de regresión.
v. Intente repetir la simulación, pero esta vez cree las condiciones para que ambos enfoques den resultados diferentes.

4. La altura... ¿se hereda?

El conjunto de datos de las alturas (`heights`) contiene las alturas (en pulgadas) de 5524 pares de madres e hijas registradas en un estudio realizado por Karl Pearson y Alice Lee en 1903.

i. Elabore un gráfico que permita ver la relación entre las alturas de las madres y las hijas. Aplique las técnicas que crea necesaria para obtener una visualización informativa y fidedigna.
ii. ¿Por qué es adecuado utilizar un modelo de regresión lineal?
iii. Ajuste el siguiente modelo de regresión lineal utilizando `{Rstan}` y *priors* que crea convenientes:

$$\begin{aligned} \text{altura hija}_i &\sim \text{Normal}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 \text{altura madre}_i \end{aligned}$$

i. Calcule la media, el desvío estándar y el intervalo de credibilidad del 95% para los parámetros del modelo utilizando el *posterior*.
ii. Interprete los coeficientes del modelo.
iii. Superponga la recta de regresión en el gráfico donde se visualiza la relación entre las variables.
iv. Obtenga el *posterior* del peso medio de una hija cuya madre mide 58 pulgadas.

5. Clima en Australia

El conjunto de datos `weather_WU` datos climáticos correspondientes a 100 días en dos ciudades de Australia: Uluru y Wollongong. Se intentará predecir la temperatura a las 3 de la tarde, utilizando otras variables.

Considere los siguientes cuatro modelos:

- m_1: `temp3pm ~ temp9am`;
- m_2: `temp3pm ~ location`;
- m_3: `temp3pm ~ temp9am + location`;
- m_4: `temp3pm ~ .`.

i. Ajuste cada uno de los modelos y construya gráficas para mostrar los parámetros obtenidos.
ii. Realice pruebas predictivas *a posteriori* para comparar los modelos.
iii. Compare los *ELPD* de cada modelo utilizando *LOO*.

6. Pingüinos

Considere el dataset de pingüinos de Palmer (`penguins`) y los siguientes modelos:

- m_1: `body_mass_g ~ flipper_length_mm`;
- m_2: `body_mass_g ~ species`;
- m_3: `body_mass_g ~ flipper_length_mm + species`;
- m_4: `body_mass_g ~ flipper_length_mm + species + flipper_length_mm:species`;
- m_5: `body_mass_g ~ flipper_length_mm + bill_length_mm + bill_depth_mm`.

i. Ajuste cada uno de los modelos y construya gráficas para mostrar los parámetros obtenidos.

ii. Realice pruebas predictivas *a posteriori* para comparar los modelos.
iii. Compare los *ELPD* de cada modelo utilizando *LOO*.

7. De tal palo...

El dataset `child_iq` contiene información de los resultados de tests de coeficiente intelectual de niños de 3 años, educación de la madre, y edad de la madre cuando dio a luz.

i. Ajuste un modelo de regresión del puntaje del bebé a los 3 años en función de la edad de la madre.
ii. Ajsute ahora un modelo que incluya la educación de la madre.
iii. Construya gráficas para mostrar los parámetros obtenidos.
iv. Realice pruebas predictivas *a posteriori* para comparar los modelos.
v. Compare los *ELPD* de cada modelo utilizando *LOO*.

8. Ingresos

El dataset `earnings` contiene los resultados de la encuesta realizada por Ross sobre Trabajo, Familia y Bienestar.

i. Ajuste un modelo que prediga ingreso en función de altura e interprete los parámetros.
ii. ¿Qué transformación sería necesaria para interpretar el intercepto como el ingreso promedio de una persona con altura promedio?
iii. Ajuste un nuevo modelo utilizando la transformación propuesta en el punto anterior y compare los *posteriors* de los coeficientes.

9. !Kung

Los !Kung son un pueblo que habita en el desierto de Kalahari entre Botsuana, Namibia y Angola. Hablan la lengua !Kung, que se destaca por su amplio uso de consonantes clic (chasquido consonántico). El !K del nombre Kung es un sonido como cuando sale un corcho de una botella.

El archivo `Howell1` contiene datos de un censo parcial realizado por Dobe Howell acerca de la población !Kung.

Considere un modelo de altura en función del peso.

i. Determine e interprete las distribuciones *a posteriori* de los parámetros.
ii. Construya un gráfico de altura en función del peso, incluya las observaciones de los individuos, la recta de regresión MAP, el intervalo del 80% para la media y y el intervalo del 80% para la altura predicha.
iii. Realice predicciones para individuos cuyos pesos son: 46.95, 43.72, 64.78, 32.59 y 54.63. Calcule la altura esperada y el intervalo del 89%.

10. Zorros urbanos

Considere del conjunto de datos sobre zorros urbanos (`foxes`). Ajuste tres modelos:

- m_1: `weight ~ area`;
- m_2: `weight ~ groupsize`;
- m_3: `weight ~ area + groupsize`.

i. Para los modelos m_1 y m_2, represente gráficamente los resultados, incluyendo la recta de regresión MAP, su intervalo del 89% y el intervalo de predicción del 89%. ¿Es alguna de las dos variables importantes para predecir la masa de un zorro?
ii. Representar gráficamente las predicciones del modelo para cada predictor, dejando el otro constante en su valor medio. ¿Qué puede decirse sobre la importancia de las variables para predecir la masa de un zorro?

11. Un *prior* informativo marca la diferencia

Considere el conjunto de datos sobre belleza y proporción de sexos (`sexratio`) . Estos datos provienen de un estudio de adolescentes estadounidenses cuyo atractivo en una escala de cinco puntos fue evaluado por entrevistadores en una encuesta cara a cara. Años más tarde, muchos de estos encuestados tuvieron hijos y se registraron ciertos atributos entre los cuales se incluyó el sexo. El objetivo del análisis es comparar la proporción de sexos de los hijos según la belleza de los padres. Para ello considere el siguiente modelo de regresión:

$$\begin{aligned} \text{pf}_i &\sim \text{Normal}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 \text{belleza}_i \end{aligned}$$

Donde pf representa la proporción de bebés de sexo femenino y belleza representa el grupo de belleza de los padres.

i. Ajuste el modelo utilizando mínimos cuadrados.
ii. Ajuste el modelo con `{RStan}` y *priors* uniformes.
iii. Compare el ajuste de ambos modelos.
iv. Explore los *priors* utilizados por `{RStan}` y la distribución predictiva *a priori*. ¿Qué puede concluir?
v. Proponga distribuciones *a priori* informativas.
vi. Ajuste el modelo utilizando `{RStan}` y los *priors* informativos.
vii. Compare el resultado con los obtenidos anteriormente y concluya.

12. ¡A la pesca de *priors*!

El conjunto de datos `fish-market` contiene mediciones morfológicas realizadas sobre pescados de diferentes especies. El objetivo es construir un modelo de regresión lineal que permita predecir el peso de los pescados en base a sus otros atributos.

Uno de los modelos propuestos es el siguiente:

$$\begin{aligned} \log(\text{Weight}_i) &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \beta_{0,j[i]} + \beta_{1,j[i]} \log(\text{Length1}_i) \end{aligned}$$

Weight_i es el peso del i-ésimo pescado en gramos y Length1_i es la longitud del i-ésimo pescado en centímetros. La letra j indexa las especies de los pescados. Por lo tanto, en este modelo cada especie tiene su propio intercepto y pendiente.

i. Implemente el modelo utilizando `{RStan}` y los siguientes *priors* no informativos:

$$\begin{aligned} \beta_{0,j[i]} &\sim \text{Normal}(0, 10) \\ \beta_{1,j[i]} &\sim \text{Normal}(0, 5) \end{aligned}$$

ii. Obtenga y visualice la distribución predictiva a *priori*.
iii. Elabore un gráfico y describa la función de densidad *a priori* de los parámetros $\beta_{0,j[i]}$ y $\beta_{1,j[i]}$.
iv. Proponga *priors* más adecuados en base a la interpretación de los parámetros del modelo y la información que tenga del problema.
v. Nuevamente, obtenga y visualice la distribución predictiva a *priori* y compare con el resultado obtenido anteriormente.
vi. Ajuste el modelo, obtenga la distribución predictiva *a posteriori* y grafíquela.

13. En búsqueda del modelo adecuado

Continuando con los datos del ejercicio anterior, El objetivo es construir un modelo de regresión lineal que permita predecir el peso de los pescados en base a sus otros atributos.

Considere los siguientes modelos

- m_1: `log(Weight) ~ 0 + Species`;
- m_2: `log(Weight) ~ 0 + Species + log(Length1)`;
- m_3: `log(Weight) ~ 0 + Species + log(Length1):Species`;
- m_4: `log(Weight) ~ 0 + Species + log(Length1):Species + log(Height)`;
- m_5: `log(Weight) ~ 0 + Species + log(Length1):Species + log(Height):Species`.

i. Ajuste cada uno de los modelos.
ii. Estime el *ELPD* de cada modelo utilizando *LOO* y seleccione el modelo más adecuado de acuerdo a este criterio.
iii. Explique el resultado.

14. Comparación de modelos

Se recopilaron datos (`mesquite`) con el fin de desarrollar un método para estimar la producción total (biomasa) de hojas de mesquite utilizando parámetros fácilmente medibles de la planta, antes de que se realice la cosecha real. Se tomaron dos conjuntos separados de mediciones, uno en un grupo de 26 arbustos de mesquite y otro en un grupo diferente de 20 arbustos de mesquite medidos en un momento diferente del año. Todos los datos se obtuvieron en la misma ubicación geográfica, pero ninguno constituyó una muestra estrictamente aleatoria. La variable de resultado es el peso total (en gramos) de material fotosintético obtenido de la cosecha real del arbusto. Las variables de entrada son:

Nombre	**Descripción**
`diam1`	Diámetro de la copa medido a lo largo del eje más largo del arbusto (metros)
`diam2`	Diámetro de la copa medido a lo largo del eje más corto (metros)
`canopy_height`	Altura de la copa
`total_height`	Altura total del arbusto
`density`	Número de tallos primarios por planta
`group`	Grupo de mediciones (0 para el primer grupo, 1 para el segundo)

i. Realice un análisis exploratorio de los datos.
ii. Ajuste el modelo `weight ~ diam1 + diam2 + canopy_height + total_height + density + group`.
iii. Explore y describa el *posterior*.
iv. Estime *ELPD* mediante *PSIS-CV* con la función `loo()`, analice los valores de las estimaciones del parámetro k de la distribución generalizada de Pareto y otros valores de la salida que crea relevant, ¿qué puede concluir?
v. Estime ELPD mediante *K-fold cross validation* con $K = 10$. Compare la estimación con el resultado obtenido mediante *PSIS-CV* y concluya.
vi. Ajuste el modelo transformando todas las variables numéricas con la función logarítmica. ¿Cómo afecta esta transformación la interpretación de los coeficientes?
vii. Estime *ELPD* mediante *PSIS-CV* con la función `loo()`. Concluya acerca de la estabilidad del cómputo. ¿Es posible comparar la la estimación con la obtenida en el inciso iv? ¿Por qué?
viii. Con ambos modelos, obtenga y grafique la distribución predictiva *a posteriori* comparándola con los datos observados ¿Cuál de los modelos representa mejor a los datos?

15. Secundarios en Portugal

Se cuenta con un conjunto de datos sobre 343 estudiantes de secundaria de Portugal (`portugal`) y se desea predecir la calificación final en matemáticas del último año en base a un gran número de predictores potencialmente relevantes.

El listado de variables se compone por: escuela del estudiante, sexo del estudiante, edad del estudiante, tipo de domicilio del estudiante, tamaño de la familia, estado de convivencia de los padres, educación de la madre, educación del padre, tiempo de viaje del hogar a la escuela, tiempo de estudio semanal, número de fracasos escolares pasados, apoyo educativo adicional, clases pagadas adicionales dentro de la materia del curso, actividades extracurriculares, si el estudiante asistió a una guardería, si el estudiante desea cursar estudios superiores, acceso a Internet en el hogar, si el estudiante tiene una relación romántica, calidad de las relaciones familiares, tiempo libre después de la escuela, si el estudiante sale con amigos, consumo de alcohol entre semana, consumo de alcohol los fines de semana, estado de salud actual y número de ausencias escolares.

Priors débilmente informativos

i. Ajuste un modelo de regresión lineal utilizando todos los predictores luego de estandarizarlos y con los siguientes *priors*:

$$\begin{aligned} \beta_k &\sim \text{Normal}(0, 2.5) \\ \sigma &\sim \text{Exponential}(1/\text{std}(y)) \end{aligned}$$

ii. Elabore un gráfico para visualizar los *posteriors* marginales y compárelos. ¿Qué puede concluir acerca de su incertidumbre?

iii. Calcule y compare la mediana del R^2 bayesiano y del R^2 calculado mediante LOO ¿Qué conclusión puede extraer de esta comparación?

iv. ¿Cuál es el número efectivo de parámetros según LOO? ¿Qué indica?

v. Obtenga muestras del *prior* y del *posterior* del R^2 bayesiano, compárelos utilizando una visualización y concluya considerando la elección de los *priors* débilmente informativos sobre β_k y σ.

Priors alternativos (I)

Si se asume que muchos predictores pueden tener poca relevancia, se pueden escalar los *priors* independientes para que la suma de la varianza de los *priors* se encuentre alrededor de un valor razonable. En este caso, se cuenta con 26 predictores y se podría suponer que la proporción de la varianza explicada por los predictores está alrededor de 0.3. Entonces, un enfoque simple consiste en asignar *priors* independientes a los coeficientes de regresión con media 0 y desviación estándar $\sqrt{0.3/26}\text{sd}(y)$ y un *prior* exponencial con media $\sqrt{0.7}\text{sd}(y)$ para σ.

i. Ajuste el modelo nuevamente utilizando los siguientes priors:

$$\begin{aligned} \beta_k &\sim \text{Normal}(0, \sqrt{\frac{0.3}{26}}\text{sd}(y)) \\ \sigma &\sim \text{Exponential}(1/\sqrt{0.7}\text{sd}(y)) \end{aligned}$$

ii. Explore la distribución *a priori* sobre R^2 y compárela con la distribución obtenida con los *priors* débilmente informativos.

iii. Calcule $ELPD$ mediante LOO y compare este modelo con el anterior.

iv. Elabore un gráfico para visualizar los *posteriors* marginales. Compare este resultado con el obtenido con los *priors* débilmente informativos.

Priors alternativos (II)

Otra alternativa es asumir que solo algunos de los predictores tienen alta relevancia y que el resto de los predictores tienen una relevancia insignificante. Una posibilidad para modelar bajo este supuesto es el *horseshoe prior* regularizado[1]. Este *prior* utiliza distribuciones normales independientes con media 0 y varianza $\tau^2\lambda_k^2$ para los coeficientes de regresión β_k y se describe a continuación:

$$\begin{aligned}
\beta_k &\sim \text{Normal}(0, \tau^2\tilde{\lambda}_k^2) \\
\tilde{\lambda}_k^2 &= \frac{c^2\lambda_k^2}{c^2 + \tau^2\lambda_k^2} \\
c &= \sqrt{c'}\text{SS} \\
c' &\sim \text{InvGamma}(0.5 \cdot SDF, 0.5 \cdot SDF) \\
\lambda_k &\sim \text{StudentT}^+(\text{df} = 1, \mu = 0, \sigma = 1) \\
\tau &\sim \text{StudentT}^+(\text{df} = 1, \mu = 0, \sigma = \text{GS})
\end{aligned}$$

con

$$\begin{aligned}
\text{GS} &= \frac{p_0}{p - p_0}\frac{\sigma}{\sqrt{n}} \\
\text{SS} &= \sqrt{\frac{0.3}{p_0}}\text{sd}(y) \\
\text{SDF} &= 4
\end{aligned}$$

donde GS, SS y SDF representan *global scale*, *slab scale* y *slab degrees of freedom*, respectivamente. Además, p representa la cantidad de predictores, 26, y p_0 la cantidad de predictores que se espera que sean relevantes.

Intuitivamente, el parámetro global τ empuja todos los β_k hacia el 0, mientras que los parámetros locales λ_k contribuyen a que algunos de los β_k escapen del 0.

i. Utilice $p_0 = 6$ para ajustar el modelo con todos los predictores y grafique y analice los *posteriors* marginales.
ii. Compare este modelo con los ajustados anteriormente en base a sus *ELPD* estimados con *LOO* y concluya.

Priors débilmente informativos con menos predictores

Ajuste el modelo de regresión con un subconjunto de predictores que crea conveniente y los *priors* débilmente informativos que se utilizaron inicialmente.

i. Visualice los *posteriors* marginales.
ii. Nuevamente, calcule y compare la mediana del R^2 bayesiano y del R^2 calculado mediante.
iii. Compare este modelo con el ajustado anteriormente en base a sus *ELPD* estimados con *LOO* y concluya sobre la capacidad predictiva de este modelo.

[1] El nombre *horseshoe* (herradura) no se relaciona con la forma de la función de densidad en sí, sino con la forma del *prior* implícito para los coeficientes de *shrinkage* (contracción) aplicados a cada parámetro.