Práctica 1.
Especificación y evaluación de argumentos causales.

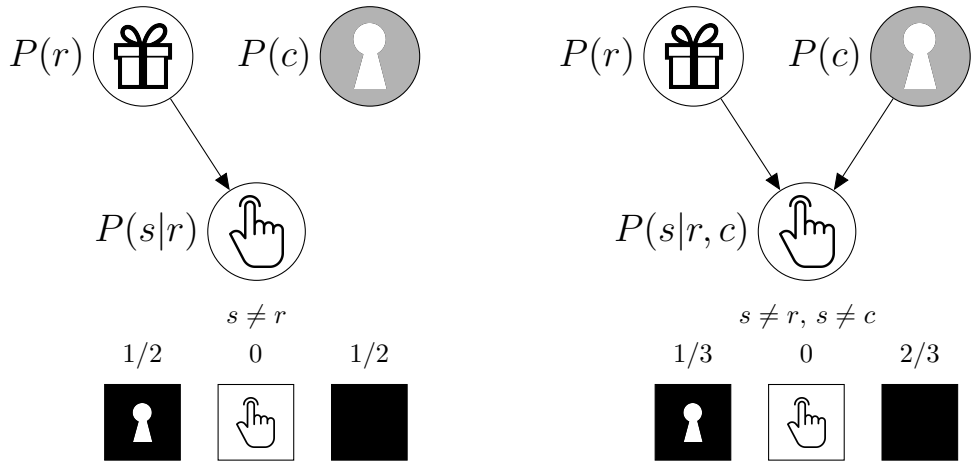
Docente: Gustavo Landfried

Inferencia Bayesiana Causal 1
2do cuatrimestre 2024
UBA - UNSAM

Índice

1. Modelo Base vs Modelo Monty Hall

En la siguiente figura se puede observar la especificación gráfica del modelo "Monty Hall" (derecha) y el modelo "Base" (izquierda) visto en la presentación. Abajo de ellos se muestra la distribución de creencias *a posteriori* sobre la posición del regalo luego de que hayamos elegido cerrar la caja 1 y nos mostraran que en la caja 2 no estaba el regalo, $P(r|s=2, c=1)$.



(a) Modelo Base (M_0)

(b) Modelo Monty Hall (M_1)

Figura 1: Modelos causales alternativos

El modelo Base (izquierda) supone que la pista s no puede señalar únicamente la caja en la que se encuentra el regalo $s \neq r$. El modelo Monty Hall (derecha) supone que la pista s no puede señalar la caja en la que se encuentra el regalo $s \neq r$ ni la caja que hemos cerrado previamente $s \neq c$. El objetivo de este ejercicio es actualizar nuestras creencias sobre los modelos causales alternativos luego de observar un conjunto datos, $P(\text{Modelo}|\text{Datos})$.

$$P(\text{Modelo}|\text{Datos}) = \frac{P(\text{Datos}|\text{Modelo})P(\text{Modelo})}{P(\text{Datos})}$$

Para ello deberemos calcular:

- La predicción que hace el modelo sobre los datos: $P(\text{Datos}|\text{Modelo})$
- La predicción de los datos realizada con la contribución de todos los modelos: $P(\text{Datos})$
- La creencia previa "honesta" sobre los modelos: $P(\text{Modelo})$

Vayamos paso a paso.

1.1. Definir la distribución de creencia conjunta como producto de las distribuciones condicionales del modelo

Las especificaciones gráficas de los argumentos causales representan dos descomposiciones alternativas de la distribución de creencias conjuntas $P(r, c, s|M)$.

$$\underbrace{P(r, c, s|M_0)}_{\substack{\text{Prior conjunto}\\ \text{hipótesis en } M_0}} = \underbrace{P(r|M_0)P(c|M_0)P(s|r, M_0)}_{\substack{\text{Relaciones causales probabilísticas}\\ \text{propuestas por el modelo } M_0}} \quad , \quad \underbrace{P(r, c, s|M_1)}_{\substack{\text{Prior conjunto}\\ \text{hipótesis en } M_1}} = \underbrace{P(r|M_1)P(c|M_1)P(s|r, c, M_1)}_{\substack{\text{Relaciones causales probabilísticas}\\ \text{propuestas por el modelo } M_1}}$$

Ambos modelos suponen que r y c son variables independientes. El modelo Base M_0, sin embargo, supone que s depende únicamente de r, $(s \neq r)$ mientras que el modelo Monty Hall M_1 considera que s depende tanto de r como de s, $(s \neq r,\ s \neq c)$.

Las distribuciones condicionales a priori sobre r y c son iguales en ambos modelos.

$$P(r|M) = P(r) = \begin{array}{c|c|c} r=0 & r=1 & r=2 \\ \hline 1/3 & 1/3 & 1/3 \end{array} \qquad P(c|M) = P(c) = \begin{array}{c|c|c} c=0 & c=1 & c=2 \\ \hline 1/3 & 1/3 & 1/3 \end{array}$$

La única diferencia entre los modelos aparece en la distribución condicional sobre la pista. En el modelo Base solo depende del regalo r.

$$P(s|r, M_0) = \begin{array}{c|ccc} & s=0 & s=1 & s=2 \\ \hline r=0 & 0 & 1/2 & 1/2 \\ r=1 & 1/2 & 0 & 1/2 \\ r=2 & 1/2 & 1/2 & 0 \end{array}$$

Y en el modelo Monty Hall depende del regalo r y la caja cerrada c. Para simplificar, mostraremos los valores cuando $c = 1$.

$$P(s|r, c=1, M_1) = \begin{array}{c|ccc} (c=0) & s=0 & s=1 & s=2 \\ \hline r=0 & 0 & 1/2 & 1/2 \\ r=1 & 0 & 0 & 1 \\ r=2 & 0 & 1 & 0 \end{array}$$

Notar que cada renglón suma 1, pues cada condicional representa una distribución de probabilidad distinta.

1.2. Mostrar que el producto de las predicciones a priori de la secuencia de datos de un episodio es igual a la probabilidad conjunta a priori

Un episodio es la secuencias completa de observaciones de todas las variables ocultas de un modelo. Al iniciar un episodio, todas las variables son ocultas (hipótesis). Al final del episodio, todas son observables. En este ejemplo, vamos a considerar que al interior de un episodio t las observaciones llegan siempre en el mismo orden: observamos primero la caja que se cierra c_t, luego la pista s_t y finalmente la posición del regalo r_t. Luego, la predicción que hace un modelo sobre esos datos puede escribirse del siguiente modo,

$$\underbrace{P(\text{Episodio} = (c_t, s_t, r_t)|\text{Modelo})}_{\substack{\text{Predicción de los datos} \\ \text{de un episodio completo}}} = \underbrace{P(c_t|\text{Modelo})}_{\substack{\text{Predicción} \\ \text{primer dato}}} \underbrace{P(s_t|c_t, \text{Modelo})}_{\substack{\text{Predicción} \\ \text{segundo dato}}} \underbrace{P(r_t|s_t, c_t, \text{Modelo})}_{\substack{\text{Predicción} \\ \text{tercer dato}}}$$

Mostrar que el producto de las predicciones a priori de la secuencia de datos de un episodio es igual a la probabilidad conjunta a priori.

$$\underbrace{P(r, c, s|\text{Modelo})}_{\substack{\text{Creencia conjunta} \\ \text{de hipótesis a priori}}} = \underbrace{P(c_t|\text{Modelo})}_{\substack{\text{Predicción} \\ \text{primer dato}}} \underbrace{P(s_t|c_t, \text{Modelo})}_{\substack{\text{Predicción} \\ \text{segundo dato}}} \underbrace{P(r_t|s_t, c_t, \text{Modelo})}_{\substack{\text{Predicción} \\ \text{tercer dato}}}$$

1.3. Simular datos con el modelo Monty Hall

Antes de evaluar los modelos necesitamos un conjunto de datos que provengan de la realidad subyacente oculta. Podríamos buscar los datos reales del programa de televisión Monty Hall y revisar si efectivamente el modelo Monty Hall propuesto es mejor que el modelo Base. Para simplificar vamos a suponer que nuestro modelo Monty Hall representa perfectamente la realidad causal subyacente y vamos a generar entonces los datos usando de nuestro propio modelo Monty Hall. Generar un conjunto de datos con $T = 16$ episodios.

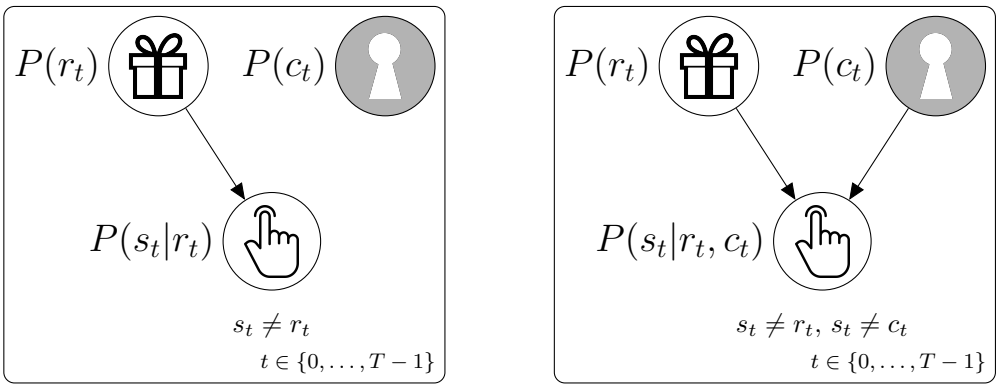
$$\text{Datos} = \{(c_0, s_0, r_0), \ldots, (c_{T-1}, s_{T-1}, r_{T-1})\}$$

1.4. Calcular la predicción a priori que hace cada uno de los modelos sobre la totalidad de la base de datos simulada

Ahora sí, podemos calcular la predicción del conjunto de datos que hace cada uno de los modelos con la contribución de todas sus hipótesis internas.

$$P(\text{Datos} = \{\underbrace{(c_0, s_0, r_0)}_{\substack{\text{Primer}\\\text{episodio}}}, \underbrace{(c_1, s_1, r_1)}_{\substack{\text{Segundo}\\\text{episodio}}}, \ldots \underbrace{(c_{T-1}, s_{T-1}, r_{T-1})}_{\substack{\text{T-ésimo}\\\text{episodio}}}\}|\text{Modelo})$$

Los modelos causales expresados en la figura 1 proponen relaciones causales probabilísticas entre las variables al interior de un episodio. En principio, estos modelos sólo están definidos para un único episodio. Para extenderlos a T episodios vamos a considerar que contamos con T repeticiones de esa misma estructura causales. Las repeticiones se especifican gráficamente mediante el uso de "placas".



(a) Modelo Base (M_0)

(b) Modelo Monty Hall (M_1)

Figura 2: Extensión de los modelos causales alternativos a T episodios mediante la notación de "placas". El subíndice t representa las repeticiones.

Es decir, entre episodios no hay flechas que vinculen las estructuras causales, por lo que ninguno modelo usa los datos de un episodio para predecir lo datos de otro episodio. Esto permite descomponer la predicción sobre el conjunto de datos sobre todos los episodios como el producto de las predicciones que lo modelos hacen al interior de cada episodio.

$$P(\text{Datos}|\text{Modelo}) = \prod_{t \in \{0,\ldots,T-1\}} P(c_t|\text{Modelo}) P(s_t|c_t, \text{Modelo}) P(r_t|s_t, c_t, \text{Modelo})$$

Calcular $P(\text{Datos}|\text{Modelo})$.

1.5. Expresar intuitivamente la diferencia de desempeño predictivo de los modelos

La predicción sobre un conjunto de datos necesariamente resulta ser un número muy cercano a 0. En este caso, en el que observamos tan solo 16 episodios, el modelo Base predijo el conjunto de datos con probabilidad $P(\text{Datos}_T|M_0) = 8{,}23e-21$ (con 21 ceros después de la coma) y el modelo Monty Hall predijo con probabilidad $P(\text{Datos}|M_1) = 3{,}37e-17$ (con 17 ceros después de la coma). Si seguimos agregando episodios al conjunto de datos, este número alcanza valores tan cercanos a cero que deja de ser posible representarlo en una computadora.

Existen dos formas alternativas de expresar este número, que son muy útiles además para ganar intuición respecto de la diferencia de desempeño entre modelos.

1.5.1. Calcular la diferencia de desempeño predictivo entre modelos expresada en órdenes de magnitud y la cantidad de creencia que el modelo Monty Hall preserva respecto del modelo Base.

Expresar la predicción en órdenes de magnitud significa hablar en términos del exponente. Por ejemplo, el exponente de las predicciones del modelo Base es alrededor de -21 y el exponente del modelo Monty Hall es alrededor de -17, como ya mencionamos anterioramente. La función que nos devuelve el exponente de un número es el logaritmo.

$$\overbrace{\log_{10} P(\text{Datos} = \{d_1, d_2, \ldots, d_N\}|M)}^{\text{Exponente de la predicción conjunta}} = \log_{10}(\overbrace{P(d_1|M)P(d_2|d_1, M)\ldots}^{P(\text{Datos}|M)}) \stackrel{*}{=} \underbrace{\log_{10} P(d_1|M)}_{\text{Exponente de } d_1} + \underbrace{\log_{10} P(d_2|d_1, M)}_{\text{Exponente de } d_2} + \ldots$$

El exponente de la predicción conjunta se puede descomponer ($\stackrel{*}{=}$) como la suma de los exponentes de las predicciones individuales. Esto permite evitar los problemas de representación computacional. Además, si calculamos la diferencia de exponentes entre los modelos obtendremos la diferencia de desempeño predictivo en órdenes de magnitud, lo que se conoce como *log Bayes Factor*.

$$\log_{10} \underbrace{\frac{P(\text{Datos}|M_1)}{P(\text{Datos}|M_0)}}_{\text{Bayes factor}} = \underbrace{\log_{10} P(\text{Datos}|M_1) - \log_{10} P(\text{Datos}|M_0)}_{\text{Diferencia predicitva en ordenes de magnitud}} \approx (-17) - (-21) = 4$$

En base 10, una diferencia de un orden de magnitud significa que uno de los modelos preservó 10 veces más creencia que el otro, dos ordenes de magnitud significa que un modelos preservó 100 veces más creencia que el otro, y así sucesivamente. Aunque estos números parezcan extraordinarios, cuatro órdenes de magnitud se considera en el límite de una diferencia no concluyente. Cuando las bases de datos crecen, la diferencia en órdenes de magnitud continuan creciendo, por lo que es normal ver diferencia de 10000, pero en órdenes de magnitud! En esos casos, para ganar intuición es útil calcular la predicción "típica".

1.5.2. Calcular la predicción típica (o media geométrica) de los modelos.

La media geométrica representa la predicción "típica"que hace un modelo de los datos. Decimos que es típica porque podemos reemplazar cada una de las predicciones que componen la secuencia por la media geometrica sin alterar el valor final. Es decir,

$$P(\text{Datos} = \{d_1, d_2, \ldots, d_N\}|M) = P(d_1|M)P(d_2|d_1, M)\cdots = \prod_{i\in\{1,\ldots,N\}} \underbrace{(P(d_1|M)P(d_2|d_1, M)\ldots)^{1/N}}_{\text{Media geométrica}}$$

Así expresada, la media geométrica tendría el mismo problema de representación computacional que señalamos al inicio de este ejercicio. Para calcularla hay que trabajar en órdenes de magnitud.

$$10^{\log_{10}(P(d_1|M)P(d_2|d_1,M)\ldots)^{1/N}} = 10^{\frac{1}{N}\left(\log_{10} P(d_1|M)+\log_{10} P(d_2|d_1,M)+\ldots\right)}$$

La expresión de la derecha puede ser calculada gracias a que es la suma de los exponentes individuales, dividido luego por N, la cantidad de datos. En nuestro caso tenemos $T = 16$ episodios, pero la cantidad total de datos es $N = T \cdot 3 = 48$, tres por cada episodio. Usaremos este número para calcular la predicción típica.

Verán que la predicción típica del modelo Base será de 0,382 y la del modelo Monty Hall será de 0,454. Dado que observamos en total $N = 48$ datos (3 datos en cada una de los $T = 16$ episodios), podemos usar la predicción típica para recuperar la predicción conjunta.

$$P(\text{Datos}|M_0) = 0{,}382^{48} \qquad\qquad P(\text{Datos}|M_1) = 0{,}454^{48}$$

En promedio (geométrico), el modelo base preserva solo el 38,2 % de la creencia previa luego de cada nueva observación, mientras que el modelo Monty Hall preserva 45,4 %. Debido a que le predicción es un proceso multiplicativo, la pérdida de creencia de los modelos es exponencial. Por eso en tan solo 48 pasos temporales el modelo Monty Hall logra preservar 4096 veces más de creencia que el modelo Base. [1]

1.6. Calcular la predicción de los datos con la contribución de todos los modelos.

Para actualizar la creencia de los modelos vamos a necesitar la probabilidad de los datos, $P(\text{Datos})$, que no es más que la predicción hecha con la contribución de todos los modelos.

$$P(\text{Datos}) \overset{\text{Regla de la suma}}{=} \sum_{\text{Modelo}} P(\text{Modelo}, \text{Datos}) \overset{\text{Regla de la producto}}{=} \sum_{\text{Modelo}} \underbrace{P(\text{Datos}|\text{Modelo})}_{\substack{\text{Predicción hecha} \\ \text{por el modelo}}} \underbrace{P(\text{Modelo})}_{\substack{\text{Creencia en} \\ \text{el modelo}}}$$

Aprovechar que tenemos la secuencia de predicciones hechas en cada uno de los episodios para calcular cómo se va actualizando la predicción conjunta a medida que se van incorporando nuevos datos.

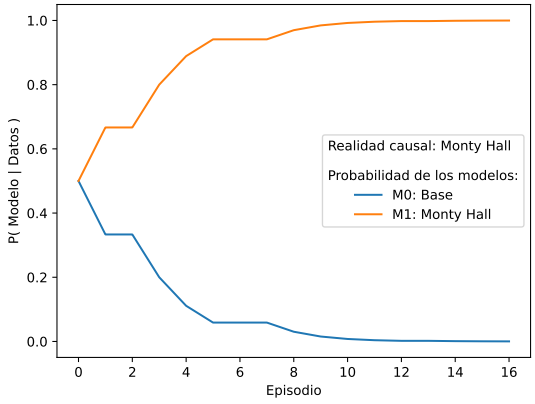
1.7. Calcular el posterior de los modelos a medida que vamos agregando datos

Ahora sí. Tenemos todos los elementos necesarios para calcular cómo se va actualizando la creencia de los modelos a medida que vamos agregando datos.

$$P(\text{Modelo}|\text{Datos}) = \frac{P(\text{Modelo}, \text{Datos})}{P(\text{Datos})}$$

1.8. Graficar el valor del posterior a medida que se observan nuevos episodios

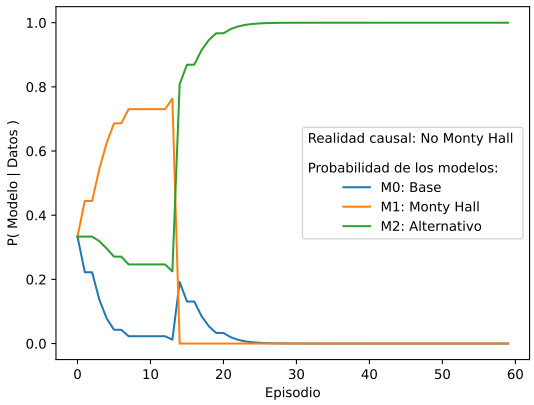
Para graficar cómo se va actualizando el posterior deberemos tener guardado el valor del posterior luego de observar cada uno de los episodios.



[1] La naturaleza exponencial de la pérdida de creencia hace que diferencias mucho menores entre las predicciones típicas de los modelos sean suficientes para identificar qué modelo funciona realmente mejor que otro. Por ejemplo, si la predicción típica del modelo Base hubiera sido 0,452, preservando apenas 0,02 % menos de creencia previa que el modelo Monty Hall, con tan solo 2000 observaciones totales encontraríamos la misma diferencia de desempeño predictivo. Aunque la diferencia de predicción típica parezca chica, si el conjunto de datos es más grande, la diferencia de desempeño predictivo también puede ser igualmente grande.

1.9. Leer los datos `NoMontyHall.csv`, proponer un modelo alternativo superior a Monty Hall y el modelo Base, y evaluarlo en función del desempeño predictivo.

Los datos del archivo `NoMontyHall.csv` contienen 2000 episodios. Los datos fueron generados con el modelo Monty Hall, salvo en algunas ocasiones en los que la persona que da la pista se olvida de tener en cuenta la caja elegida, y actúa siguiendo el modelo Base. Crear un modelo alternativo que tenga un desempeño similar al modelo Monty Hall. Calcular la secuencia de predicciones de los tres modelos en esta nueva base de datos (Modelo Base, Monty Hall y Alternativo). Calcular la diferencia de desempeño predictivo en órden de magnitud entre los modelos. Calcular la media geométrica de las predicciones de los diferentes modelos. Graficar el posterior de los tres modelos en los primeros 60 episodios. Debería quedar algo similar a lo siguiente.



Como ayuda, en la siguiente figura mostramos la especificación del modelo alternativo a Base y Monty Hall utilizado. En este modelo se agregan dos variable. La primera es una variable de "acción" (a_t) que indica el comportamiento adquirido por la persona que da la pista en el episodio t.

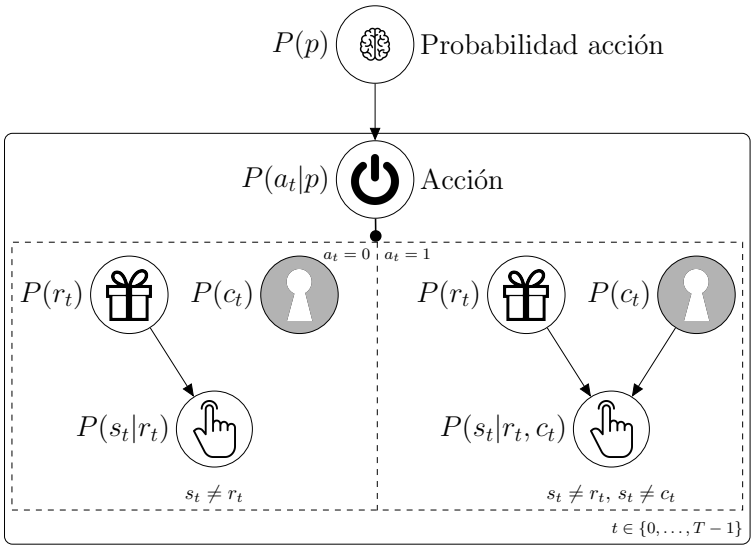


Figura 5: Modelo alternativo al Base y MontyHall. Dependiendo de si la persona se acuerda o no, actúa en función de un modelo u otro.

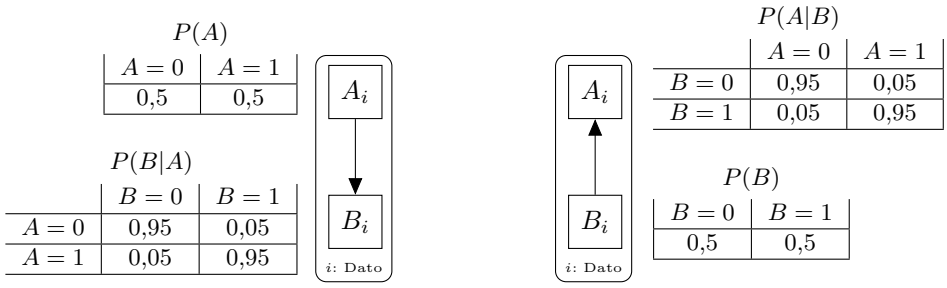
La probabilidad conjunta es

$$P(r, c, s, a, p|M_2) = P(r)P(c)P(s|r, M_0)^{1-a}P(s|r, c, M_1)^a P(a|p)P(p)$$

En este caso, cuando queremos hacer la predicción al interior de un episodio t vamos a tener dos variables ocultas, a_t y p que deberemos integrar usando la regla de la suma.

2. Modelos AcausaB vs Modelo BcausaA

Hasta ahora pudimos usar las reglas de las probabilidad para evaluar modelos causales alternativos. En este ejercicios intentaremos descubrir la dirección de una relación causal entre dos variables A y B.



(a) Modelo AcausaB

(b) Modelo BcausaA

Figura 6: Incertidumbre sobre la dirección de la relación causal

2.1. Generar datos con el modelo AcausaB

Vamos a suponer que el modelo AcausaB representa perfectamente la realidad causal subyacente. Generar un conjunto de datos con $T = 16$ episodios.

$$\text{Datos} = \{(a_0, b_0), \ldots, (a_{15}, b_{15})\}$$

2.2. Evaluar el desempeño predictivo de los modelos causales alternativos sobre los datos sintéticos generados en el punto anterior

$$P(\text{Datos}|\text{Modelo}) = \prod_t P(\text{Episodio}_t = (a_t, b_t)|\text{Modelo})$$

2.3. Actualizar la creencia respecto de los modelos causales alternativos luego de ver los datos.

Calcular el posterior de los modelos dado los datos.

$$P(\text{Modelo}|\text{Datos}) = \frac{P(\text{Datos}|\text{Modelo})P(\text{Modelo})}{P(\text{Datos})}$$

2.4. Dar sus conclusiones.

Explicar el posterior de los modelos obtenido.

3. Modelos polinomiales de complejidad creciente.

La gran mayoría de los modelos de inteligencia artificial, incluyendo las redes neuronales, se construyen a partir de modelos que postulan relaciones "lineales" entre las hipótesis. El modelo más básico es la famosa regresión lineal. Dado un conjunto de datos $\{(x_1, y_1), \ldots, (x_T, y_T)\}$, un modelo causal lineal afirma que el valor de la variable y_i se generar en función de x_i y un conjunto de hipótesis ocultas h, tal que

$$y \leftarrow h_0 + h_1 \cdot x$$

h_0 representa el valor base de y cuando x es neutral ($x = 0$), y h_1 representa el efecto causal que la varibale x tiene sobre y, el cual es proporcional a su propio valor ($h_1 \cdot x$), dando una relación lineal entre el valor de la causa x y el valor de su efecto y. De modo similar, podemos construir relaciones más complejas, como son los polinomios de grado M.

$$y \leftarrow h_0 + h_1 \cdot x + h_2 \cdot x^2 + \cdots + h_M \cdot x^M = \sum_{i=0}^{M} h_i \cdot x^i$$

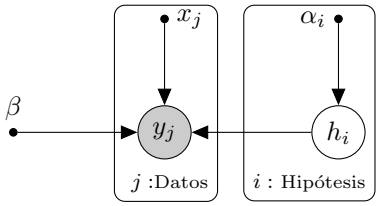
Es interesante notar que un polinomio de grado $M - 1$ es un caso especial de un polinomio de grado M en el que el $h_M = 0$. Es decir, cuánto más complejo sea el polinomio, más flexibilidad tiene el modelo para representar la relación entre x e y. Debido a que los datos se miden con un error que se produce con desvío estándar β centrado en cero, la probabilidad condicional de observar un dato es,

$$p(y|x, \boldsymbol{h}, \text{Modelo} = M) = \mathcal{N}\left(y \,\middle|\, \sum_{i=0}^{M} h_i \cdot x^i,\ \beta^2\right)$$

Como tenemos incertidumbre respecto de los valores de las hipótesis h_i, proponemos una distribución que creencias *a priori* alrededor del cero.

$$p(h_i) = \mathcal{N}(h_i \,|\, 0, \alpha_i^2)$$

Luego, la especificación gráfica del modelo lineal es



Quisiéramos actualizar nuestra creencia respecto de las hipótesis h al interior de cada modelo y actualizar nuestra creencia sobre los modelos causales alternativos. Ninguno de estas dos objetivos se pudo realizar de forma exacta hasta las vísperas del siglo 21. En este ejercicio nos interesa comparar los resultados que se obtienen mediante los métodos propuestos durante el siglo 20, basados en estimadores puntuales, respecto de los resultado que se obtienen de evaluar todo el espacio de hipótesis (y modelos) mediante la aplicación estricta de las reglas de la probabilidad. En el primer caso usaremos el método OLS (por *Ordinary Least Squares*) del paquete `statsmodels` y para el segundo caso usaremos nuestra propia implementación disponible en el archivo `ModeloLineal.py`[2].

[2]La derivación matemática de la regresión lineal la pueden encontrar en el capítulo 2 del libro de Bishop PRML. El presente ejercicio está extraído del capítulo 3 del mismo libro.

3.1. Generar 20 datos alrededor de una período de una sinoidal

Supongamos que los datos se generan de una realidad causal subyacente completamente distinta, siguiendo una función sinoidal en el rango $x \in [-\frac{1}{2}, \frac{1}{2}]$.

$$p(x) = \text{Unif}(x| -0{,}5, 0{,}5)$$
$$p(y|x) = \mathcal{N}(y|\texttt{sin}(2\pi x), \beta^2)$$

Al graficar la función objetivo (línea) y los datos (puntos) se soberva lo siguiente.

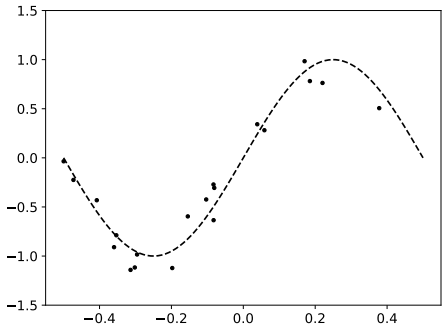


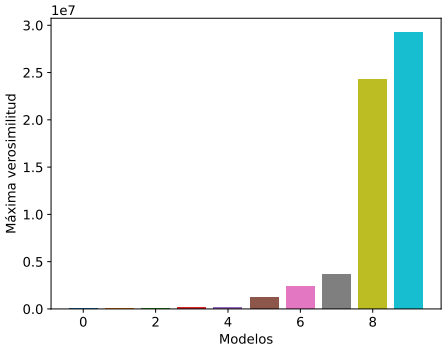
Figura 7: Datos generados de la "función objetivo" (realidad causal subyacente)

3.2. Graficar el valor de máxima verosimilitud obtenido por los modelos polinomiales de grado 0 a 9

Debido a la complejidad computacional de la aplicación estricta de las reglas de la probabilidad, durante el siglo 20 se propusieron una gran cantidad de criterios arbitrarios para la selección de una única hipótesis del espacio, como es el principio de máxima verosimilitud.

$$\arg\max_h \; p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{h}, M_M) \overset{*}{=} \arg\min_h \sum_{j\in\{1,\dots,N\}} (y_j - \sum_i^M h_i \cdot x_j^i)^2$$

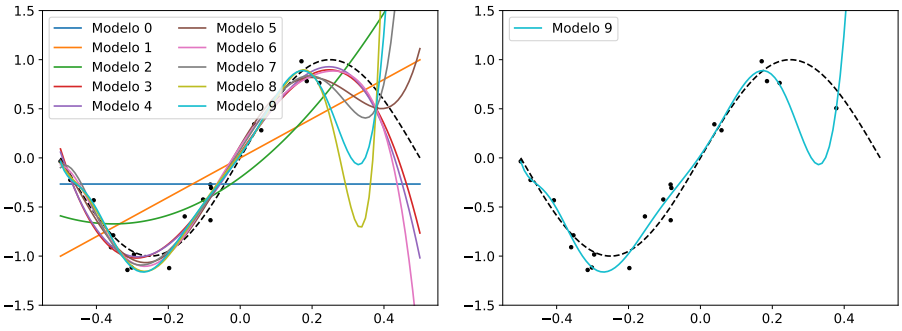
Por propiedades de este tipo de modelos ($\overset{*}{=}$) encontrar las hipótesis que mejor predicen es lo mismo que minimizar la suma de las distancias cuadradas entre el verdadero valor y el valor medio propuesto. Al graficar el valor de máxima verosimilitud obtenido con cada uno de los modelos veremos que a medida que aumentamos la complejidad de los modelos aumenta también el valor de máxima verosimilitud.



Esto ocurre debido a que cuanto más flexible es un modelo, más se puede acercar a todos los puntos. Por lo tanto, si usamos el criterio de máxima verosimilitud para seleccionar modelo, elegiríamos siempre el modelo más complejo.

3.3. Graficar las curvas obtenidas con cada modelo mediante máxima verosimilitud.

Seleccionar el modelo de mayor complejidad no es deseable. Para ver por qué, vamos a graficar las curvas obtenidas por cada uno de los modelos de grado 0 a 9.



(a) Todos los modelos (b) Modelo de máxima verosimilitud

Figura 9: Ajuste de los polinomios por máxima verosimilitud.

En la figura 9a podemos ver que los modelos de grado 0 a 2 no tienen la complejidad suficiente para acercarse a los datos como lo hacen los modelo de complejidad 3 o superior. En la figura 9b destacamos la curva que se obtiene con el modelo más complejo, de grado 9.

A diferencia del modelo 3 (color rojo en figura 9a) que se mantiene cerca de la función objetivo, el modelo de grado 9 se separa significativamente de la función objetivo alrededor de los valores $x = 0{,}3$ y $x = 0{,}5$. La excesiva flexibilidad hace que los modelos más complejos adopten formas extrañas que suelen alejarse más de la función objetivo oculta, por lo que la predicción sobre nuevos datos del modelo 3 termina siendo superior al del modelo 9 que habíamos elegido por máxima verosimilitud.

Esto se conoce se conocen en el área de inteligencia artificial con el nombre de sobreajuste o *overfitting* y veremos que es efecto secundario de utilizar métodos arbitrarios de selección de hipótesis (cómo máxima verosimilitud) y no una propiedad indeseable del sistema de razonamiento probabilístico.

3.4. Evaluación de la predicción "en línea" que hacen los modelos ajustados por máxima verosimilitud.

Para evaluar correctamente los modelos alternativos deberíamos calcular la evidencia de los modelos ($P(\text{Datos}|\text{Modelo})$), que no es más que la predicción *a priori* que hace el modelo de todos los datos, lo que es lo mismo que el producto de las predicciones individuales del siguiente dato usando los datos ya observados como información previa.

$$P(\text{Datos} = \{d_1, d_2, \dots\}|M_D) = P(d_1|M_d)P(d_2|d_1, M_D)\dots$$

Este producto de predicciones representa lo que nos ocurriría efectivamente si usáramos el modelo en la vida real: ajustamos el modelo con los datos que ya tenemos disponibles (entrenamiento) y evaluamos el desempeño en los datos nuevos (testeo). Para simular este proceso, vamos a implementar un procedimiento en el cual predecimos los datos individuales ajustando los modelos por máxima verosimilitud sobre los datos observados previamente.

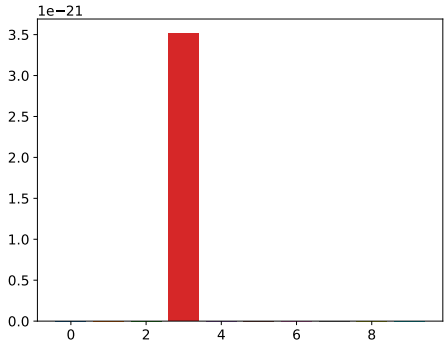


Figura 10: Producto de las predicciones a prior de los modelos ajustados por máxima verosimilitud.

Si evaluáramos los modelos en base al desempeño predictivo, dado por el producto de las predicciones a periori, rechazaríamos todos los modelos salvo el de grado 3. Esto también es un caso de sobreajuste (overfitting), pues en caso de recibir datos por fuera del período, tendremos bajo desempeño predictivo debido a que el modelo de grado 3 no tendrá la flexibilidad suficiente para acercarse a la función objetivo. El problema del sobreajuste (overfitting) ocurre como efecto secundario de seleccionar una única hipótesis del espacio y no poder computar de forma exacta la evidencia de los modelos, $P(\text{Datos}|\text{Modelo})$.

Cuando seleccionamos de forma arbitraria una única hipótesis no podemos computar la evidencia de los modelos, $P(\text{Datos}|\text{Modelo})$, debido a que las predicciones que hacen los modelos debería realizarse con la contribución de todas las hipótesis, y no con una única hipótesis seleccionada previamente.

$$P(y_i|\underbrace{x_1, y_1}_{d_1}, \ldots, \underbrace{x_{i-1}, y_{i-1}}_{d_{i-1}}, x_i, M_D) = \int_{\boldsymbol{h}} P(y_i|d_1, \ldots, d_{i-1}, x_i, \boldsymbol{h}, M_D) P(\boldsymbol{h}|d_1, \ldots, d_{i-1}, M_D)$$

$$\approx P(y_i|d_1, \ldots, d_{i-1}, x_i, \underbrace{\arg\max_{\boldsymbol{h}} P(d_1, \ldots, d_{i-1}|\boldsymbol{h}, M_D)}_{\text{Hipóteis de máxima verosimilitud}}, M_D)$$

Es decir, aproximamos las predicciones que deberían hacer los modelos con la contribución de todas las hipótesis mediante la predicción que obtenemos con la hipótesis que maximizaba la verosimilitud en el paso anterior.

3.5. Más criterios arbitrarios de selección de hipótesis: los regularizadores

Para evitar algunos de los efectos secundarios de la selección arbitraria de hipótesis mediante máxima verosimilitud, se han propuesto una gran variedad de otros criterios arbitrarios de selección. Una familia alternativa de criterios muy difundida son los llamados regularizadores. En vez de seleccionar la hipótesis que mejor predice los datos (máxima verosimilitud) elegimos la hipótesis que tienen mayor creencia luego de observar los datos (máximo a posteriori).

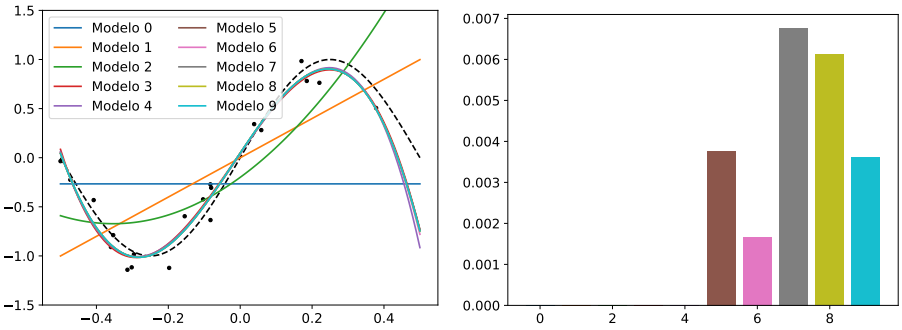
$$\arg\max_h \underbrace{p(\boldsymbol{h}|\overbrace{\boldsymbol{y}, \boldsymbol{x}}^{\text{Datos}}, M_M)}_{\text{Posterior}} = \arg\max_h \overbrace{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{h}, M_M)}^{\text{Verosimilitud}} \overbrace{p(\boldsymbol{h}|M_M)}^{\text{Prior}}$$

$$\overset{*}{=} \arg\min_h \underbrace{\sum_j (y_j - \sum_i h_i \cdot x_j^i)^2}_{\text{Distancia entre dato y predicción media}} + \underbrace{\frac{\alpha}{2}||h||^2}_{\text{Penalización}}$$

Por propiedades del modelo ($\overset{*}{=}$) maximizar el producto de la verosimilitud con el prior es equivalente a minimizar la suma de distancias cuadradas entre el dato y la predicción, más la

suma de la distancia cuadrada de las hipótesis al origen pesada por la precisión a priori sobre las hipótesis α. Esto hace que la hipótesis que se seleccione no sea la que más se acerca a los datos, sino la que mejor balance tenga entre la distancia y la penalización.

Veamos cómo se desempeñan los modelos bajo este nuevo criterio arbitrario de selección de hipótesis. Vamos a proceder del modo similar a lo realizado en el ejercicio anterior. Para calcular el posterior exacto sobre las hipótesis vamos a usar nuestra propia implementación disponible en el archivo `ModeloLineal.py`. Debido a que la penalización depende del prior, vamos a calcular el posterior usando un prior 100 veces más informativo que el prior que definimos al inicio del ejercicio.



(a) Ajuste por Máximo a posteriori (b) Predicciones a priori "en línea"

Figura 11: Selección arbitraria de hipótesis por máximo a posterior con prior informativo.

El prior informativo obliga a que los valores de las hipótesis no puedan alejarse de 0. Esto hace que los modelos más complejos pierdan flexibilidad y se comporten como si fueran modelos más simples. De esta forma, se logra mejorar el desempeño de los modelos complejos incluso cuando evaluamos sus predicciones a priori, entrenando el modelo en línea cada vez que recibe un nuevo dato.

En este caso estamos rechazando todos los modelos simples, quedándonos con los modelos de mayor complejidad. Si bien esto podría parecer una solución al problema del ejercicio anterior, no lo es. Así como el rechazo de todos los modelos salvo el de grado 3 hacía que no pudiéramos predecir potenciales datos futuros que estuvieran fuera del período observado, la selección de los modelos más complejos no ocurre por mérito de su complejidad sino por la reducción de flexibilidad que le impusimos mediante la regularización. En términos prácticos, los modelos complejos están funcionando como una variante apenas más flexible que el modelo 3 regularizado. La regularización de hipótesis entonces no resuelve el problema, pues cuando veamos datos por fuera del período observado, estos modelos tampoco van a ser capaces de predecir los nuevos datos.

$$P(\text{Datos}) = \sum_{m \in \{5,\ldots,9\}} P(\text{Datos}|\text{Modelo} = m)P(\text{Modelo} = m) \approx P(\text{Datos}|\text{Modelo} = 3)$$

3.6. El balance natural de las reglas de la probabilidad.

Para evaluar correctamente las hipótesis y modelos causales alternativos es suficiente con actualizar las creencias aplicando estrictamente las reglas de la probabilidad. Para eso necesitamos calcular la probabilidad a posteriori de los los modelos dado los datos.

$$P(\text{Modelo} = m|\text{Datos}) = \frac{\overbrace{P(\text{Datos}|\text{Modelo} = m)}^{\text{Evidencia}} P(\text{Modelo} = m)}{\sum_i P(\text{Datos}|\text{Modelo} = i)P(\text{Modelo} = i)}$$

Para calcular de forma exacta la evidencia del modelo (en órdenes de magnitud) podemos usar el método `lm.log_evidence()` de nuestro paquete `ModeloLineal.py`. Suponiendo que no tenemos

preferencia a priori por ningún modelo, vamos a usar la evidencia para calcular el posterior exacto de los modelos.

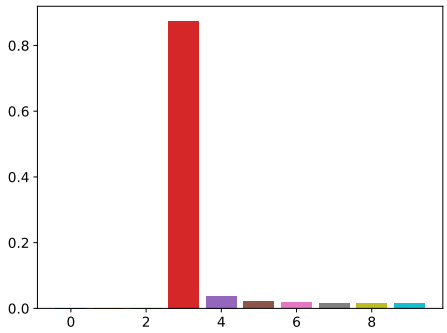


Figura 12: Posterior exacto de los modelos.

Cuando evaluamos correctamente las hipótesis, encontramos que: el modelo que tiene la menor complejidad necesaria (grado 3) es el que obtiene mayor creencia a posteriori; los modelos que tienen menor complejidad de la necesaria (grado 0 a 2) son rechazados; y los modelos que tienen mayor complejidad de la necesaria (grado 4 a 9) tienen baja probabilidad, pero no son rechazados. A diferencia de lo que ocurría cuando seleccionábamos hipótesis de forma arbitraria, evaluando correctamente las hipótesis garantizamos nuestra capacidad para predecir potenciales datos que aparecieran en el futuro por fuera del del período observado gracias a que tenemos disponibles todavía modelos más complejos capaces de explicarlos.

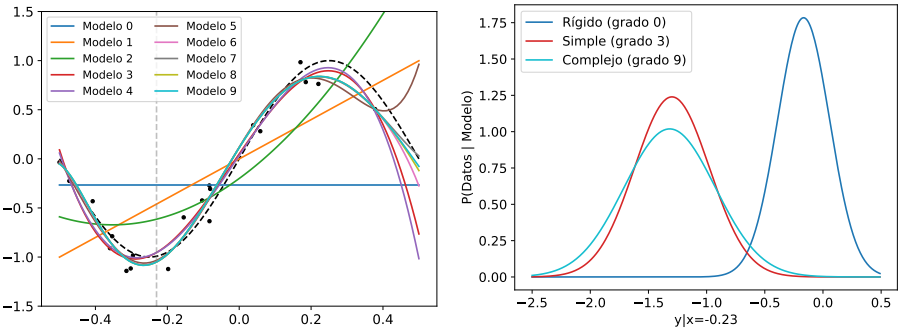
$$P(\text{Datos}) = \sum_m P(\text{Datos}|\text{Modelo} = m)P(\text{Modelo} = m)$$

En definitva, los problemas que emergen cuando se seleccionan las hipótesis mediante criterios arbitrarios simplemente no forman parte del sistema de razonamiento en contextos de incertidumbre. Por eso, la aplicación estricta de las reglas de la probabilidad, siguiendo el principio de "no mentir", es el razonamiento óptimo en contextos de incertidumbre.

3.7. Cómo se explica el balance natural de las reglas de la probabilidad

En probabilidad las predicciones son distribuciones de probabilidad que deben integrar 1. Esto produce una balance natural entre los modelos debido a que ningún modelo es superior a otro en términos absolutos. Todos los modelos tienen una región en donde ganan y todos tienen una región en donde pierden. Veamos esto de forma concreta.

En la figura 13a graficamos la media del posterior de los modelos basado en un prior no informativo. No estamos graficando la incertidumbre de los modelos, que representaría la predicción que hacen los modelos de los datos. Para graficar esa incertidumbre vamos a hacer un corte en la figura 13a cuando x vale $-0{,}23$ (linea vertical).



(a) Media del posterior con priori no informativo

(b) Predicción en $x = -0{,}23$

Figura 13: El balance natural entre modelos

En la figura 13b hacemos un corte en línea punteada ($x = -0{,}23$) y graficamos la predicción que hace el modelo más rígido (grado 0), el modelo simple (grado 3) y el modelo más complejo (grado 9) luego de ver el cuarto dato, vamos a observar tres distribuciones gaussianas con diferente media y diferente desvío estándar. Podemos ver que el modelo más rígidos (grado 0) concentra su creencia en una región lejana al valor verdadero. Los modelos simples (grado 3) y complejo (grado 9) distribuyen su creencia alrededor del verdadero valor pero de forma distinta: cuanto mayor flexibilidad tienen los modelos, mayor tendencia a distribuir la creencia en una región más amplia de valores posibles. En términos un poco más esquemáticos (con $x = 0{,}1$) lo que ocurre es.

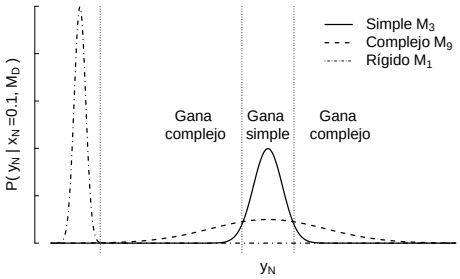


Figura 14: Esquema de la predicción de los modelos

De esta forma, el modelo rígido se rechaza por su incapacidad de llevar su creencia a la región correcta. Pero entre el modelo simple y el modelo complejo hay una balance en el cual existen ciertas regiones donde el modelo simple gana y ciertas regiones donde el modelo más complejo gana.

4. Efecto causal del sexo biológico sobre la altura.

Vamos a utilizar un conjunto de datos sobre alturas de un grupo de personas y 3 variables adicionales: sexo biológico, contextura de la madre, y altura de la madre. En este ejemplo vamos a proponer modelos causales alternativos entre estas variables.

4.1. Abrir el archivo `alturas.csv` y visualizar los datos

El archivo de datos tiene la siguiente estructura,.

```
   id  altura sexo contextura_madre altura_madre
0   1   172.7    M         mediana        159.8
1   2   171.5    M         mediana        160.3
```

```
2   3   162.6   F         mediana       160.5
3   4   174.1   M         mediana       159.8
4   5   168.3   M         mediana       158.3
```

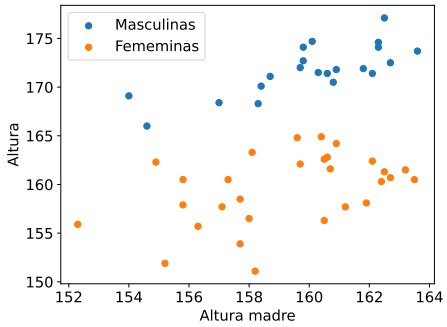


Figura 15: Datos de alturas

4.2. Definir 3 modelos causales alternativos

Si bien los datos son de dudosa procedencia, vamos a jugar un juego de inferencia causal. Vamos a proponer y evaluar tres modelos causales alternativos. En el modelo base vamos a suponer que la altura de la madre tienen un efecto causal lineal sobre la altura de su descendencia.

$$\text{Modelo Base: } \texttt{altura} = h_0 + h_1 \cdot \texttt{altura\_madre}$$

En el modelo biológico vamos a suponer que el sexo tiene un efecto causal adicional sobre la altura.

$$\text{Modelo Biológico: } \texttt{altura} = h_0 + h_1 \cdot \texttt{altura\_madre} + h_2 \cdot \mathbb{I}(\texttt{sexo} = F)$$

Cuando el sexo es masculino, la función identidad $\mathbb{I}(\texttt{sexo} = M)$ vale 1 y en caso contrario vale 0, lo que prende y apaga la hipótesis h_2, que reprersentar el efecto causal adicional del sexo biológico.

Por último vamos a plantear el modelo identitario, en el que vamos a suponer que la identidad de la persona (y no el sexo) tiene efecto causal sobre la altura.

$$\text{Modelo grupos al azar: } \texttt{altura} = h_0 + h_1 \cdot \texttt{altura\_madre} + h_2(\texttt{ID mod max(ID)/2})$$

4.3. Computar la evidencia de los modelos causales alternativos

En la siguiente figura graficamos los órdenes de magnitud de evidencia ($P(\text{Datos}|\text{Modelo})$) de los 3 modelos causales alternativos.

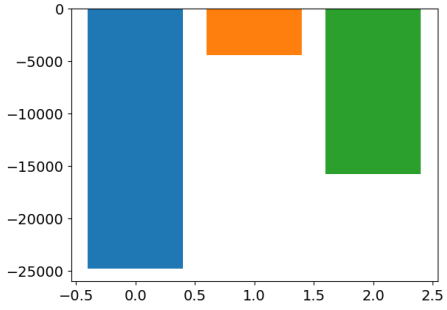


Figura 16: Evidencia en escala logarítmica de los modelos causales alternativos

4.4. Computar la media geométrica de los modelos causales alternativos

La media geométrica

$$P(\text{Datos} = \{d_1, d_2, \ldots, d_N\}|M) = P(d_1|M)P(d_2|d_1, M)\cdots = \prod_{i\in\{1,\ldots,N\}} \underbrace{(P(d_1|M)P(d_2|d_1, M)\ldots)^{1/N}}_{\text{Media geométrica}}$$

4.5. Computar el posterior de los modelos

$$P(\text{Modelo}|\text{Datos}) = \frac{P(\text{Datos}|\text{Modelo})P(\text{Modelo})}{P(\text{Datos})}$$