Estadística Bayesiana
Parte III

Irving Gómez Méndez

Validación del Modelo

Simulación de la predictiva previa

Note que también se puede simular una muestra de la predictiva previa, para ello basta con simular una muestra de los parámetros $\theta_1, \ldots, \theta_m$, a partir de la distribución previa y luego simular $Y_j \sim P(Y|\theta_j)$.

Poder simular de la predictiva previa puede ayudar a discriminar entre distintas distribuciones previas. Pues al simular de la predictiva previa podemos observar las consecuencias de las distribución previa sobre la variable de interés Y. Muchas de las técnicas convencionales para decidir la distribución previa pueden dar lugar a distribuciones más bien absurdas.

Validación del modelo

Una vez que hemos completado los dos primeros pasos del análisis bayesiano: contrucción de un modelo de probabilidad y cómputo de la distribución posterior, no deberíamos ignorar el paso (relativamente sencillo) de validar qué tan bien se ajusta nuestro modelo a los datos y a nuestro conocimiento del fenómeno.

Dado que de antemano sabemos que nuestro modelo no puede incluir todos los aspectos de la realidad, podemos averiguar cuáles aspectos no son capturados por el modelo. Y, sobre todo sobre la plausibilidad de nuestro modelo para el propósito para el que fue construido. **No se trata de preguntarnos si nuestro modelo es falso o verdadero, sino de cuáles son las principales deficiencias de nuestro modelo.**

Estadísticas de prueba y p-valor frecuentista

Desde un enfoque frecuentista se define una estadística de prueba $T(\mathbf{Y})$ como una estadística de los datos usada para comparar los datos observados contra una réplica generada bajo el modelo propuesto. De esta manera, el p-valor clásico se define como

$$p_C = \mathbb{P}(T(\mathbf{Y}^{rep}) \geq T(\mathbf{Y})|\theta).$$

Esto es, p_C es la probabilidad de obtener una estadística más extremista que la observada, con θ fijo. En donde θ puede ser un valor 'nulo' en pruebas de hipótesis o alguna estimación como el estimador de máxima verosimilitud.

Contexto bayesiano

En el contexto bayesiano, aprovechamos que contamos con una distribución generativa dada por la distribución predictiva. Si el modelo ajusta bien, entonces **datos replicados con nuestro modelo generativo deberían parecer similares a los datos observados**. Dicho de otra manera, los datos observados deben de parecer plausibles al considerar la distribución predictiva posterior. Nuestra técnica básica para validar el ajuste de un modelo es simular datos de la distribución predictiva posterior y compararlos con los datos observados.

Significancia estadística y significancia práctica

El objetivo no es responder la pregunta: ¿los datos provienen del modelo supuesto?, cuya respuesta casi siempre será no, sino cuantificar las discrepancias entre los datos y el modelo, y saber cuándo esas discrepancias podrían haber surgido por puro azar, bajo los supuestos del modelo.

Si la falla del modelo en un aspecto importante es grande, podemos pensar en cambiar el modelo, si no, podemos ignorarla si no afecta las conclusiones principales. **El p-valor mide 'significancia estadística' no 'significancia práctica'.**

Réplicas de los datos

Para que la comparación sea correcta los datos replicados, $\mathbf{Y}^{rep}$ deben de ser (cómo su nombre indica) réplicas de los datos observados. Es decir, $\mathbf{Y}^{rep}$ debe de ser de la misma dimensión que $\mathbf{Y}$, y si nuestro modelo cuenta con variables predictoras $\mathbf{X}$, entonces debemos usar exactamente los mismos valores de las variables predictoras.

Cantidad de prueba y p-valor bayesiano

Para medir la discrepancia entre el modelo ajustado y los datos, definimos una cantidad de prueba $T(\mathbf{Y}, \theta)$. Al ser θ una variable aleatoria, la cantidad de prueba puede depender no sólo de los datos sino también del valor de θ.

El p-valor bayesiano se define entonces como la probabilidad de que la cantidad de prueba evaluada en los datos replicados, $T(\mathbf{Y}^{rep}, \theta)$, sea más extrema que la cantidad de prueba evaluada en los datos observados.

$$\begin{aligned} p_B &= \mathbb{P}(T(\mathbf{Y}^{rep}, \theta) \geq T(\mathbf{Y}, \theta) | \mathbf{Y}) \\ &= \mathbb{E}\left[\mathbb{1}_{T(\mathbf{Y}^{rep}, \theta) \geq T(\mathbf{Y}, \theta)} | \mathbf{Y} \right] \end{aligned}$$

Estimación del p-valor bayesiano

Como $p_B = \mathbb{E}\left[\mathbb{1}_{T(\mathbf{Y}^{rep},\theta)\geq T(\mathbf{Y},\theta)}|\mathbf{Y}\right]$, entonces una manera de estimar el p-valor es simulando m valor de la distribución posterior $\theta_1,\ldots,\theta_m$ y m réplicas, $\mathbf{Y}_1^{rep},\ldots,\mathbf{Y}_m^{rep}$ de la distribución predictiva posterior y calcular

$$\hat{p}_B = \frac{1}{m}\sum_{j=1}^{m}\mathbb{1}_{T(\mathbf{Y}_j^{rep},\theta_j)\geq T(\mathbf{Y},\theta_j)}.$$

Validación marginal del modelo

Además, podemos calcular la distribución marginal predictiva $p(Y_i|\mathbf{Y})$ y compararla con los datos para buscar datos atípicos o validar el modelo de manera individual, es decir, podemos calcular

$$p_i = \mathbb{P}(T(Y_i^{rep}, \theta) \geq T(Y_i, \theta)|\mathbf{Y}).$$

Si Y_i es continua, podemos hacer $T(Y_i, \theta) = Y_i$. Si Y_i es discreta, una opción es

$$p_i = \mathbb{P}(Y_i^{rep} > Y_i|\mathbf{Y}) + \frac{1}{2}\mathbb{P}(Y_i^{rep} = Y_i).$$

Si los p-valores marginales se concentran cerca de 0 y 1, entonces los datos están sobredispersos comparados con el modelo, mientras que si los p-valores se concentran en 0.5 los datos tienen menor dispersión que la que estima el modelo.

- Mostrar ejemplo de la estatura de la población !Kung.
- Mostrar ejemplo con los datos de Newcomb.
- Mostrar ejemplo de independencia en lanzamientos Bernoulli.

Regresión Bayesiana

Regresión de la estatura en el peso

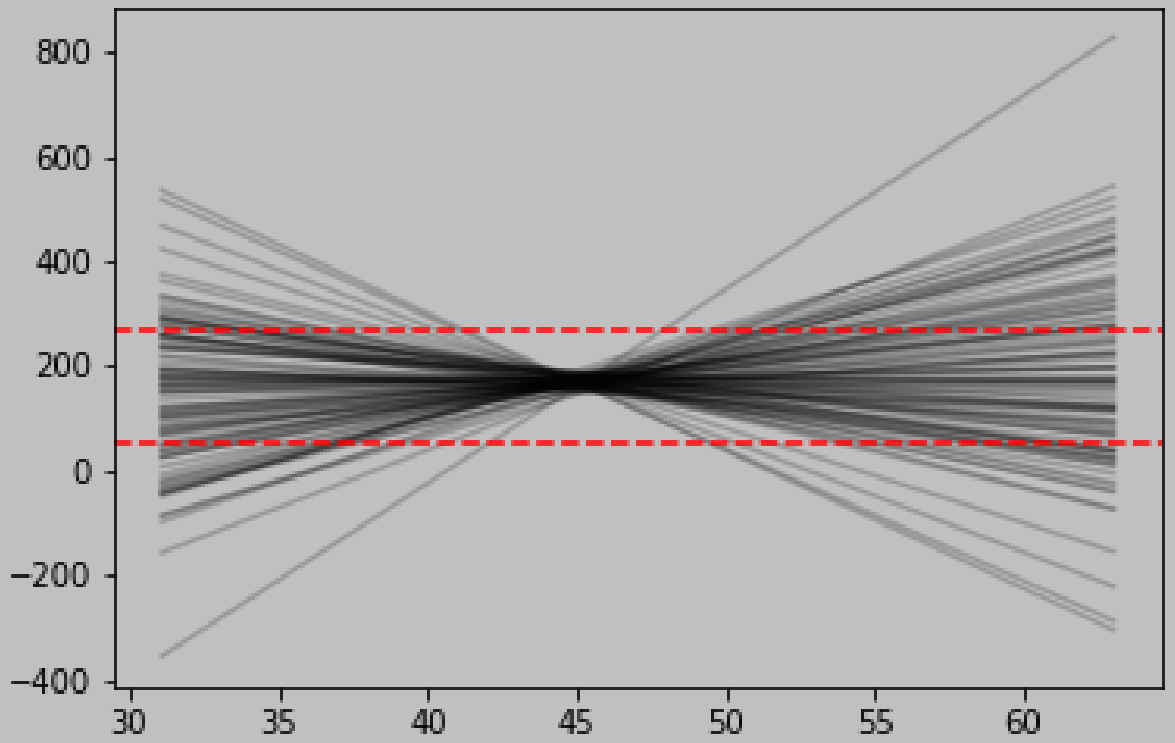
Para estos ejercicios vamos a usar los datos de la comunidad !Kung. Consideraremos sólo a los adultos, esto porque la estatura está fuertemente relacionada con la edad hasta antes de la adultez.

Sea h_i la estatura del i-ésimo individuo, w_i su peso y $\bar{w}$ el peso promedio de los adultos. Vamos a considerar el modelo:

$$
\begin{aligned}
h_i|\mu_i, \sigma &\sim \mathcal{N}(\mu_i, \sigma^2) \\
\mu_i &= \alpha + \beta(w_i - \bar{w}) \\
\alpha &\sim \mathcal{N}(170, 10^2) \\
\beta &\sim \mathcal{N}(0, 10^2) \\
\sigma &\sim \mathcal{U}(0, 33)
\end{aligned}
$$

Note cuando $w_i = \bar{w}$, $\mu_i = \alpha$. Es decir cuando el peso es igual al promedio de la población la estatura promedio será α, así que hace mucho sentido modelar $\alpha \sim \mathcal{N}(170, 10^2)$. Para σ simplemente usamos una distribución con poca información. Pero **¿por qué hemos puesto esa previa para β?**

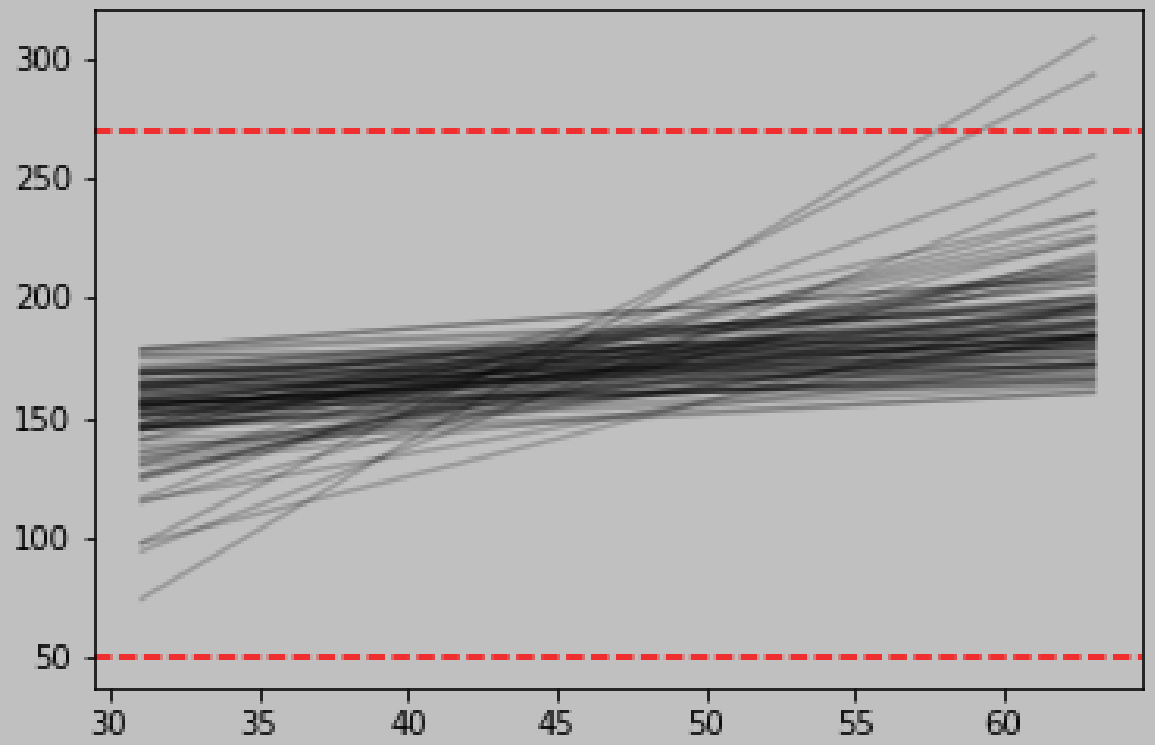
Una manera de verificar qué tan sensatas son estas previa es simulando valores de α y β, $\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m$ y luego graficar $\alpha_j + \beta_j(w - \bar{w}), \quad j = 1, \ldots, m.$



¡Evidentemente esta previa para β es absurda!

Cualquiera sabe que el peso y la estatura guardan (hasta cierto punto) una correlación positiva, por lo que es sensato considerar una distribución previa estrictamente positiva para β. Por lo que consideramos el siguiente modelo:

$$
\begin{aligned}
h_i|\mu_i, \sigma &\sim \mathcal{N}(\mu_i, \sigma^2) \\
\mu_i &= \alpha + \beta(w_i - \bar{w}) \\
\alpha &\sim \mathcal{N}(170, 10^2) \\
\beta &\sim \text{lognormal}(0, 1) \\
\sigma &\sim \mathcal{U}(0, 33)
\end{aligned}
$$


300
250
200
150
100
50
30
35
40
45
50
55
60

Estimador MAP

Una vez establecido el modelo podemos crear una rendija de valores para α, β y σ. Y como antes, simular una muestra de la posterior. Con esta muestra podemos estimar puntualmente los parámetros y graficar la recta con mayor probabilildad a posteriori $\hat{\alpha} + \hat{\beta}(w - \bar{w})$.

Intervalos de confianza

Como μ depende de los parámetros y estos tienen una distribución posterior, entonces μ también cuenta con una distribución posterior. En particular fijando un peso w, y contando con una muestra a posteriori $\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m$ podemos obtener una muestra a posteriori $\mu_1, \ldots, \mu_m$ usando la relación:

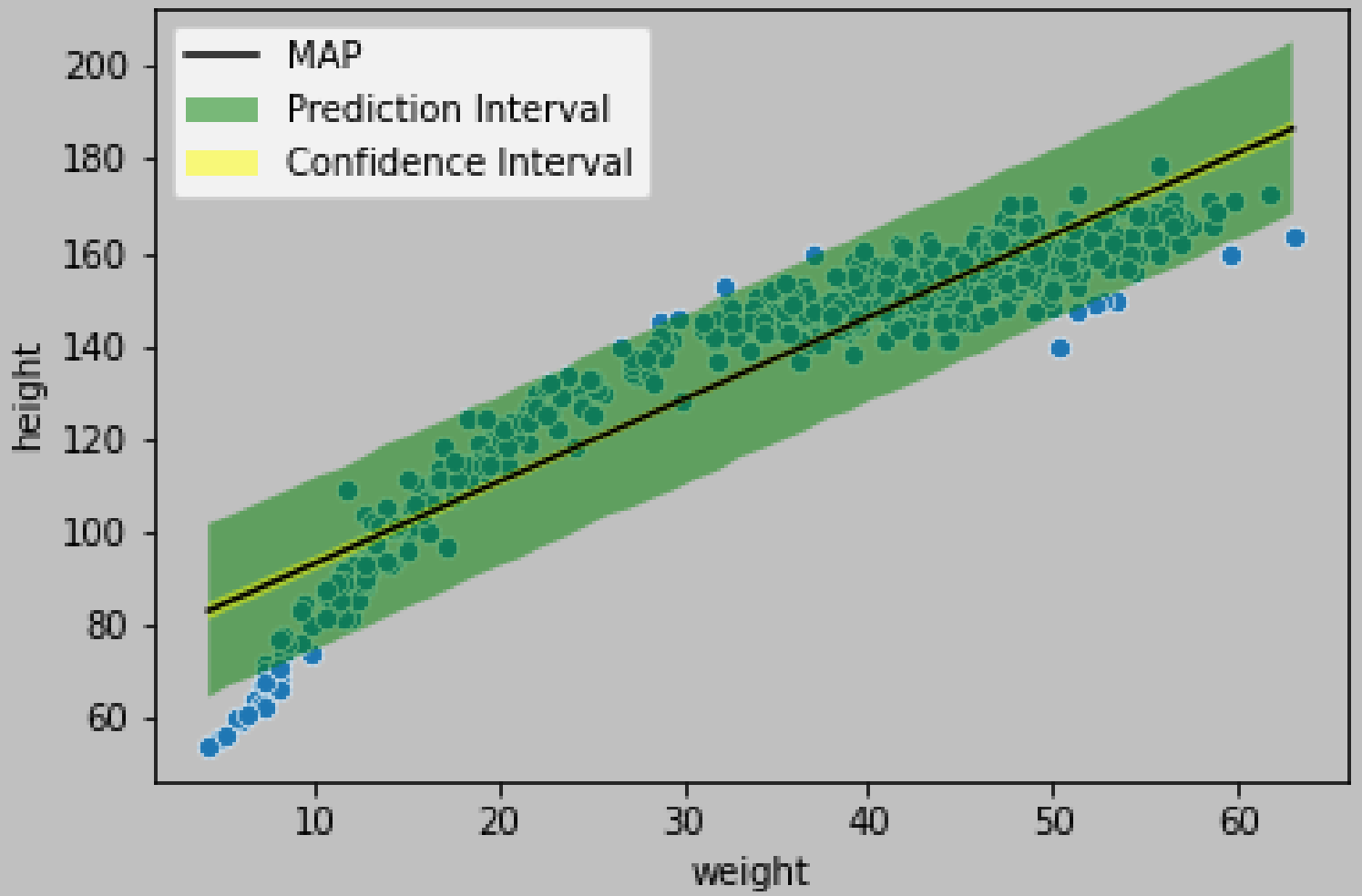
$$\mu_j = \alpha_j + \beta_j(w - \bar{w}).$$

A partir de esta muestra podemos obtener un intervalo de confianza para la estatura media dado el peso w. Variando w desde el valor más bajo hasta el valor más alto podemos obtener un intervalo de confianza para la función de regresión.

Intervalos de predicción

Finalmente, para agregar intervalos de predicción hay que recordar que $h \sim \mathcal{N}(\mu, \sigma^2)$. Así para un peso fijo w podemos obtener una muestra de la posterior de μ como se explicó antes, $\mu_1, \ldots, \mu_m$. Además podemos simular una muestra de la posterior de σ, $\sigma_1, \ldots, \sigma_m$. Por lo tanto podemos obtener una muestra de la posterior de h, $h_1, \ldots, h_m$, donde $h_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$. A partir de la muestra $h_1, \ldots, h_m$ podemos obtener un intervalo de predicción para el peso w. Variando w desde el valor más bajo hasta el valor más alto podemos obtener un intervalo de predicción para la función de regresión.

- Mostrar ejemplo de regresión de la estatura de la población !Kung.


MAP
Prediction Interval
Confidence Interval
height
weight
200
180
160
140
120
100
80
60
10
20
30
40
50
60

Aproximación por cuadratura

Es claro que una de las limitantes del método de aproximación usando una rendija es que escala pobremente cuando incrementamos el número de parámetros en el modelo (maldición de la alta dimensionalidad). De los ejemplos anteriores también hemos visto que las posteriores de los parámetros tienden a tomar una forma acampanada cuando el tamaño de la muestra es grande (teorema del límite central).

Esta observación nos permite introducir el método de aproximación por cuadratura. Hay que recordar que la densidad normal tiene esta forma de campana y que el logaritmo es proporcional a $\lambda(Y - \mu)^2$, i.e. es una función cuadrática (una parábola). Usando este hecho podemos aproximar la posterior de los parámetros mediante una distribución normal.

Regresión de la estatura en el peso

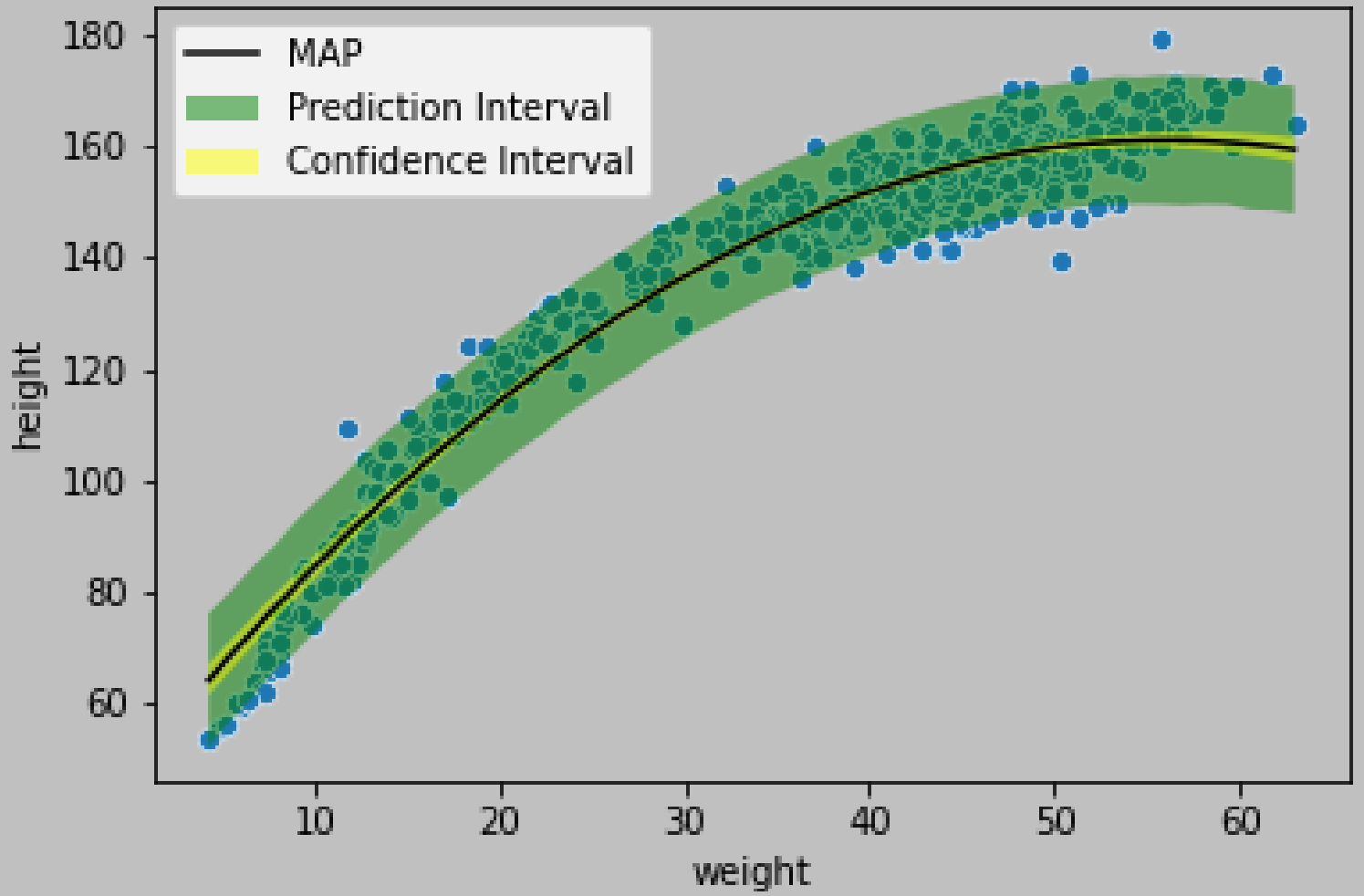
Consideremos una vez más el ejemplo de predecir la estatura de la población !Kung según su peso, pero ahora tomemos todos los individuos de la muestra. Ahora estamos dispuestos a aceptar que pueda haber mayor variabilidad en la estatura así que cambiamos la distribución previa de σ por $\sigma \sim \mathcal{U}(0, 50)$. Antes de ajustar el modelo estandarizamos la altura y el peso (restamos la media y dividimos por el desvío padrón), más adelante veremos por qué hemos hecho esto. Para estimar la posterior de los parámetros usamos la aproximación por cuadratura.

- Mostrar ejemplos de aproximación por cuadratura.

El modelo no es satisfactorio por lo que en su lugar proponemos el siguiente modelo:

$$
\begin{aligned}
h_i|\mu_i, \sigma &\sim \mathcal{N}(\mu_i, \sigma^2) \\
\mu_i &= \alpha + \beta_1 y_i + \beta_2 y_i^2 \\
\alpha &\sim \mathcal{N}(170, 10^2) \\
\beta_1 &\sim \text{lognormal}(0, 1) \\
\beta_2 &\sim \mathcal{N}(0, 1) \\
\sigma &\sim \mathcal{U}(0, 50),
\end{aligned}
$$

donde $y_i = (w_i - \bar{w})/s_w$.
Como estamos considerando una potencia de la variable predictora conviene estandarizarla antes para evitar posibles problemas numéricos.


MAP
Prediction Interval
Confidence Interval
height
weight
180
160
140
120
100
80
60
10
20
30
40
50
60

Tarea 5

- Dejar tarea 5.

Modelos Jerárquicos

Bioassay example

Rat tumors

The analysis using the data to estimate the prior parameters, which is called empirical Bayes, can be viewed as an approcimation to the complete hierarchical Bayesian analysis. In its simplest form the parameters θ_j form an independent sample from a prior (or population) distribution governed by some unknown parameter vector ϕ; thus,

$$p(\theta|\phi) = \prod_{j=1}^{J} p(\theta_j|\phi).$$

In general ϕ is unknown and thus has its own (hyper)prior distribution $p(\phi)$.

The joint prior distribution is

$$p(\theta, \phi) = p(\theta|\phi)p(\phi),$$

and the joint posterior distribution is

$$\begin{aligned} p(\theta, \phi|\mathbf{Y}) &\propto p(\theta, \phi)p(\mathbf{Y}|\theta, \phi) \\ &= p(\theta, \phi)p(\mathbf{Y}|\theta) \\ &= p(\theta|\phi)p(\phi)p(\mathbf{Y}|\theta). \end{aligned}$$

In order to create a joint probability distribution for (θ, ϕ), we must assign a prior distribution to ϕ. But we must be careful when using an improper prior density and check that the resulting posterior distribution is proper.

Deriving the conditional posterior for θ might be easy, since

$$\begin{aligned} p(\theta|\phi, \mathbf{Y}) &\propto p(\theta|\phi)p(\mathbf{Y}|\theta) \\ &= f(\phi, \mathbf{Y})p(\theta|\phi)p(\mathbf{Y}|\theta). \end{aligned}$$

where $f(\phi, \mathbf{Y})$ is the 'constant' normalizing factor that depends on ϕ and $\mathbf{Y}$.
For the posterior of ϕ, we can calculated it integrating the joint posterior distribution over θ:

$$p(\phi|\mathbf{Y}) = \int p(\theta, \phi|\mathbf{Y})d\theta.$$

Another way to calculate $p(\phi|\mathbf{Y})$ is using the probability conditional formula:

$$p(\phi|\mathbf{Y}) = \frac{p(\theta, \phi|\mathbf{Y})}{p(\theta|\phi, \mathbf{Y})}$$

$$\propto \frac{p(\theta|\phi)p(\phi)p(\mathbf{Y}|\theta)}{f(\phi, \mathbf{Y})p(\theta|\phi)p(\mathbf{Y}|\theta)}$$

$$\propto \frac{p(\phi)}{f(\phi, \mathbf{Y})}$$

For the rat tumor experiments, we assumed

$$Y_j|\theta_j \sim \mathsf{Binomial}(n_j, \theta_j),$$

with the number of rats, n_j, known. And

$$\theta_j \sim \mathsf{Beta}(\alpha, \beta).$$

If we wish a prior distribution for (α, β), we must check that the posterior distribution is proper.

$$p(\theta, \alpha, \beta|\mathbf{Y}) \propto p(\alpha, \beta)p(\theta|\alpha, \beta)p(\mathbf{Y}|\theta)$$

$$\propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1} \prod_{j=1}^{J} \theta_j^{Y_j}(1-\theta_j)^{1-Y_j}$$

Then,

$$p(\theta|\alpha,\beta,\mathbf{Y}) \propto \prod_{j=1}^{J} \theta_j^{\alpha+Y_j-1}(1-\theta_j)^{\beta+n_j-Y_j-1}.$$

That is, given (α,β), the components of θ have independent beta posterior densities,

$$\theta_j|\alpha,\beta,\mathbf{Y} \sim \mathsf{Beta}(\alpha+Y_j,\beta+n_j-Y_j),$$

and

$$p(\theta|\alpha,\beta,\mathbf{Y}) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+Y_j)\Gamma(\beta+n_j-Y_j)}\theta_j^{\alpha+Y_j-1}(1-\theta_j)^{\beta+n_j-Y_j-1}$$

$$p(\theta|\alpha,\beta,\mathbf{Y}) = f(\alpha,\beta,\mathbf{Y})p(\theta|\alpha,\beta)p(\mathbf{Y}|\theta),$$

where

$$f(\alpha,\beta,\mathbf{Y}) = \prod_{j=1}^{J} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+Y_j)\Gamma(\beta+n_j-Y_j)} \left[\binom{n_j}{Y_j}\right]^{-1}.$$

Thus,

$$p(\alpha,\beta|\mathbf{Y}) \propto p(\alpha,\beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+Y_j)\Gamma(\beta+n_j-Y_j)}{\Gamma(\alpha+\beta+n_j)}.$$

A 'noninformative' prior that yields a proper hyperprior posterior (see BDA3 page 110) is given by $p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$. Note that we could avoid the mathematical effort of checking the integrability of the posterior density if we were to use a proper hyperprior distribution.

Hierarchical normal model

Checar las notas hechas y el ejemplo de las notas.

Experiment in 8 schools

Evaluación y Comparación de Modelos

Sean p y q las distribuciones predictivas posteriores bajo dos modelos distintos $\mathcal{P}$ y $\mathcal{Q}$, y suponga que la verdadera densidad de nuestros datos es f. Vamos a preferir el modelo $\mathcal{P}$ sobre el modelo $\mathcal{Q}$ si la divergencia KL entre f y p es menor que la divergencia KL entre f y q.

Es decir, preferimos $\mathcal{P}$ sobre $\mathcal{Q}$ si

$$KL(f||p) < KL(f||q)$$

$$\begin{aligned}\Leftrightarrow &\mathbb{E}_{Y\sim f}[\log f(Y)] - \mathbb{E}_{Y\sim f}[\log p(Y|\mathbf{Y})] \\ &< \mathbb{E}_{Y\sim f}[\log f(Y)] - \mathbb{E}_{Y\sim f}[\log q(Y|\mathbf{Y})]\end{aligned}$$

$$\Leftrightarrow \mathbb{E}_{Y\sim f}[\log p(Y|\mathbf{Y})] > \mathbb{E}_{Y\sim f}[\log q(Y|\mathbf{Y})]$$

Sin embargo, todavía tenemos el problema de tener que calcular un valor esperado con respecto a la verdadera distribución f. Para solucionar este problema podemos hacer uso una vez más de los datos $Y_1, \ldots, Y_n$ que sabemos que tienen densidad f.

Por lo tanto, preferiremos el modelo $\mathcal{P}$ sobre el modelo $\mathcal{Q}$ si

$$\frac{1}{n}\sum_{i=1}^{n}\log p(Y_i|\mathbf{Y}) > \frac{1}{n}\sum_{i=1}^{n}\log q(Y_i|\mathbf{Y})$$
$$\Leftrightarrow \sum_{i=1}^{n}\log p(Y_i|\mathbf{Y}) > \sum_{i=1}^{n}\log q(Y_i|\mathbf{Y})$$

Por otro lado, recuerde que la distribución predictiva posterior puede ser escrita de la siguiente manera

$$\begin{aligned} p(Y|\mathbf{Y}) &= \int_{\Theta} p(Y, \theta|\mathbf{Y}) d\theta \\ &= \int_{\Theta} p(Y|\theta) p(\theta|\mathbf{Y}) d\theta, \end{aligned}$$

por lo tanto, preferiremos el modelo asociado a p sobre el modelo asociado a q si

$$\sum_{i=1}^{n} \log \int_{\Theta} p(Y_i|\theta) p(\theta|\mathbf{Y}) d\theta > \sum_{i=1}^{n} \log \int_{\Theta} q(Y_i|\theta) q(\theta|\mathbf{Y}) d\theta$$

A

$$\sum_{i=1}^{n} \log \int_{\Theta} p(Y_i|\theta)p(\theta|\mathbf{Y})d\theta$$

se le llama *log pointwise predictive density* (lppd).

Si contamos con una muestra $\theta_1, \ldots, \theta_S$ de la distribución posterior de θ, entonces podemos aproximar $\int_{\Theta} p(Y|\theta)p(\theta|\mathbf{Y})d\theta$ de la siguiente manera

$$\int_{\Theta} p(Y|\theta)p(\theta|\mathbf{Y})d\theta \approx \frac{1}{S}\sum_{s=1}^{S} p(Y|\theta_s)$$

Por lo tanto, preferiremos el modelo $\mathcal{P}$ sobre el modelo $\mathcal{Q}$ si

$$\sum_{i=1}^{n} \log \left[\frac{1}{S} \sum_{s=1}^{S} p(Y_i|\theta_s) \right] > \sum_{i=1}^{n} \log \left[\frac{1}{S} \sum_{s=1}^{S} q(Y|\theta_s) \right].$$

A

$$\sum_{i=1}^{n} \log \left[\frac{1}{S} \sum_{s=1}^{S} p(Y_i|\theta_s) \right]$$

la llamaremos *computed log pointwise predictive density* (computed lppd).

Note que si S es lo suficientemente grande entonces computed lppd aproximará muy bien a lppd. Por esta razón algunos autores definen el lppd como esta segunda expresión.

Aunque, en principio necesitamos calcular la densidad de cada observación Y_i en cada muestra del parámetro θ_s, en la práctica podemos enfrentar problemas numéricos, por lo que es preferible reescribir el *computed lppd* de la siguiente manera:

$$\begin{aligned}
\text{computed lppd} &= \sum_{i=1}^{n} \log \left[\frac{1}{S} \sum_{s=1}^{S} p(Y_i|\theta_s) \right] \\
&= \sum_{i=1}^{n} \left[\log \sum_{s=1}^{S} p(Y_i|\theta_s) - \log(S) \right] \\
&= \sum_{i=1}^{n} \left[\log \sum_{s=1}^{S} \exp \left\{ \log p(Y_i|\theta_s) \right\} - \log(S) \right]
\end{aligned}$$

La función $\log \sum \exp \{\cdot\}$ suele estar programada en varios lenguajes de programación de manera eficiente, manteniendo la precisión númerica, normalmente se le llama logsumexp. Así, podemos calcular el *computed lppd* como:

$$\text{computed lppd} = \sum_{i=1}^{n} [\text{logsumexp}(\log p(Y_i|\theta_s)) - \log(S)]$$

El caso ideal para evaluar un modelo sería contar con una muestra de entrenamiento, con la que se ajusta el modelo; y una muestra de prueba, con la que se calcula la métrica de desempeño del modelo.

Es decir, suponga que además de nuestra muestra de entrenamiento $Y_1, \ldots, Y_n$, también contamos con una muestra de prueba $\tilde{Y}_1, \ldots, \tilde{Y}_m$, con la que calcularíamos nuestra métrica

$$\text{computed lppd} = \sum_{i=1}^{m} \Big[\text{logsumexp}(\log p(\tilde{Y}_i|\theta_s)) - \log(S)\Big]$$

Sin embargo, cuando no contamos con una muestra de prueba, debemos tener en cuenta que estamos usando dos veces nuestros datos. Una primera vez para ajustar el modelo y obtener una muestra de la posterior, y una segunda vez para evaluar el modelo. Esto provocará una evaluación más optimista del modelo.

El plan es entonces calcular el lppd y después agregar algún término de penalización que corrija la estimación y evitar así elegir modelos sobreajustados.

Criterios de Información

Por razones históricas a las medidas de exactitud predictiva se les llama criterios de información, y típicamente se definen en términos de la devianza. Es importante aclarar que no existe un único concenso al definir los criterios de información y, por lo tanto, sus definiciones pueden variar ligeramente.

Devianza

Típicamente se define la devianza como -2 veces la logverosimilitud de los datos fijando los parámetros en algún valor, es decir $-2 \log p(\mathbf{Y}|\hat{\theta})$. Aunque desde un enfoque puramente bayesiano, algunos autores la definen como -2 veces el lppd. El -2 es simplemente por razones históricas.

Criterio de la información de Akaike (AIC)

El criterio de información más conocido es el criterio de la información de Akaike (AIC). La corrección más sencilla está basada en el comportamiento asintótico normal de la distribución posterior.

Sea k el número de parámetros estimados en el modelo. En el caso límite (o en casos especiales como cuando se cuenta con un modelo normal lineal con varianza conocida y previa uniforme), restar k de la verosimilitud dado el estimador de máxima verosimilitud es la manera de corregir la sobreestimación del poder predictivo del modelo

$$\log p(\mathbf{Y}|\hat{\theta}_{mle}) - k.$$

El AIC se define como la expresión anterior multiplicada por -2,

$$\text{AIC} = -2 \log p(\mathbf{Y}|\hat{\theta}_{mle}) + 2k.$$

Desde el enfoque bayesiano, algunos autores redefinen el AIC como

$$\text{AIC} = -2\mathsf{lppd} + 2k.$$

Cuando ya no se cuenta con un modelo lineal con previas uniformes se vuelve inapropiado simplemente agregar el número de parámetros. Para modelos con previas informativas o con una estructura jerárquica, el número efectivo de parámetros depende de la varianza de los parámetros al nivel del grupo.

Criterio de información de la devianza (DIC)

El DIC es una especie de versión bayesiana del AIC, haciendo dos cambios. El primero es reemplazando el estimador ded máxima verosimilitud por la media a posteriori $\hat{\theta}_{Bayes} = \mathbb{E}(\theta|\mathbf{Y})$ y el segundo es que reemplaza a k por el número efectivo de parámetros p_{DIC}.

El criterio de la información de la devianza se define entonces como

$$\text{DIC} = 2\log p(\mathbf{Y}|\hat{\theta}_{\text{Bayes}}) + 2p_{\text{DIC}}.$$

Existen al menos dos maneras distintas de definir el número efectivo de parámetros:

$$p_{\text{DIC}} = 2\left(\log p(y|\hat{\theta}_{\text{Bayes}}) - \mathbb{E}_{\theta\sim p(\theta|\mathbf{Y})}(\log p(\mathbf{Y}|\theta))\right),$$

usando una muestra de la posterior $\theta_1, \ldots, \theta_S$, esta expresión puede ser aproximada por

$$\text{computed}p_{\text{DIC}} = 2\left(\log p(y|\hat{\theta}_{\text{Bayes}}) - \frac{1}{S}\sum_{s=1}^{S}(\log p(\mathbf{Y}|\theta_s))\right).$$

Si la media posterior se encuentra lejos del máximo a posteriori, existe el problema de que p_{DIC} tome un valor negativo.

Una versión alternativa para el número efectivo de parámetros está dada por:

$$p_{\text{DIC alt}} = 2\mathbb{V}_{\theta\sim p(\theta|\mathbf{Y})}\left(\log p(\mathbf{Y}|\theta)\right),$$

usando una muestra de la posterior $\theta_1, \ldots, \theta_S$, esta expresión puede ser aproximada por

$$\mathsf{computed}p_{\text{DIC alt}} = 2\mathbb{V}_{s=1}^{S} \log p(\mathbf{Y}|\theta_s),$$

donde $\mathbb{V}_{s=1}^{S}$ representa la varianza muestral

$$\mathbb{V}_{s=1}^{S} a_s = \frac{1}{S-1}\sum_{s=1}^{S}(a_s - \bar{a})^2$$

Criterio de la información de Watanabe-Akaike (WAIC)

El criterio de la información de Watanabe-Akaike o *widely applicable information criterion* (WAIC) también acepta al menos dos maneras distintas de definir el número efectivo de parámetros

$$p_{\text{WAIC 1}} = 2\sum_{i=1}^{n}\left(\log(\mathbb{E}_{\theta\sim p(\theta|\mathbf{Y})}p(Y_i|\theta)) - \mathbb{E}_{\theta\sim p(\theta|\mathbf{Y})}(\log p(Y_i|\theta))\right).$$

$$\text{computed}p_{\text{WAIC 1}} = 2\sum_{i=1}^{n}\left(\log\left(\frac{1}{S}\sum_{s=1}^{S}p(Y_i|\theta_s)\right) - \frac{1}{S}\sum_{s=1}^{S}\log p(Y_i|\theta_s)\right).$$

La segunda manera de definir el número efectivo de parámetros está dada por

$$p_{\text{WAIC 2}} = \sum_{i=1}^{n} \mathbb{V}_{\theta \sim p(\theta|\mathbf{Y})} \log p(Y_i|\theta_s),$$

$$\text{computed} p_{\text{WAIC 2}} = \sum_{i=1}^{n} \mathbb{V}_{s=1}^{S} \log p(Y_i|\theta_s).$$

Así, se define el WAIC como

$$\text{WAIC} = -2\text{lppd} + 2p_{\text{WAIC}}.$$

Validación cruzada

También se puede corregir el optimismo generado al usar dos veces los datos aplicando validación cruzada. Sin embargo, validación cruzada puede ser computacionalmente intensiva, pues requiere varias particiones de los datos. En un caso extremo *leave-one-out cross-validation* (LOO-CV) requiere n particiones de los datos.

Las técnicas basadas en validación cruzada no serán discutidas aquí.