EST-46115: Modelación Bayesiana

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Comparación de modelos.
Objetivo: Ya hemos visto cómo diagnosticar y criticar nuestros modelos bayesianos de manera interna. Estudiaremos mecanismos de validación y comparación de modelos que nos servirán para llevar un proceso iterativo de construcción y crítica de modelos en conjunto.
Lectura recomendada: Parte de la discusión se ha tomado de [3], el Capítulo 2.5 de [6], el Capítulo 7 de [7], y el Capítulo 7 de [2].

1. INTRODUCCIÓN

Hemos estudiado usar evaluación y crítica de modelos por medio distribuciones predictivas. Esto nos permitió criticar un modelo en aislamiento. Es decir, sólo considerando el modelo –la elección de verosimilitud y previa– que estamos utilizando. Por otro lado, en la sección pasada estudiamos mecanismos basados en generación de datos sintéticos para evaluar si el modelo está bien implementado.

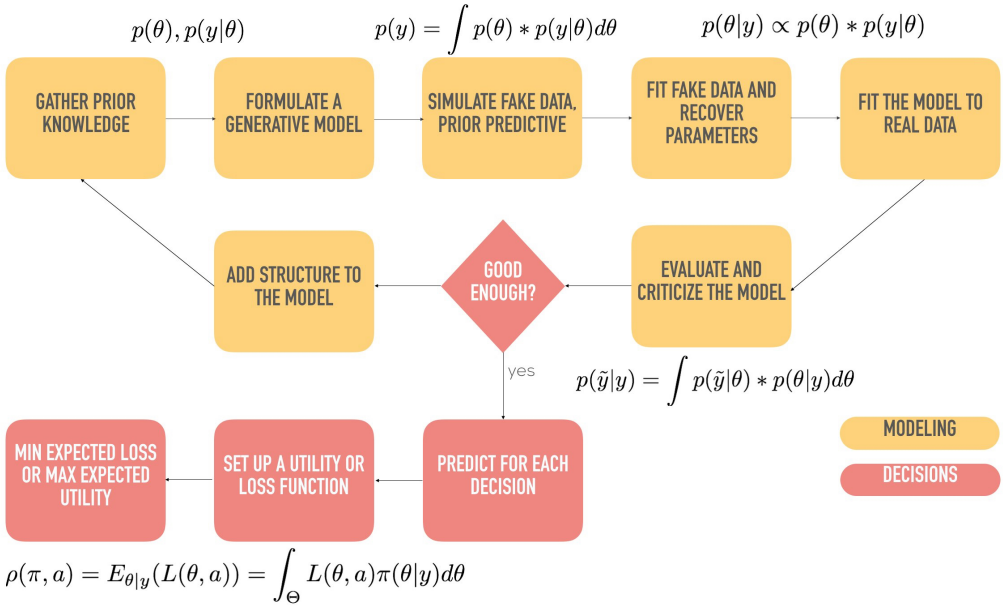


FIGURA 1. *Flujo de trabajo bayesiano. En esta sección nos concentraremos en realizar comparaciones de modelos.*

2. COMPARACIÓN DE MODELOS PROBABILISTICOS

En el contexto de modelos probabilísticos tiene sentido preguntarnos no sólo por la capacidad predictiva del modelo (en términos puntuales) sino también por la confianza del modelo en dichas predicciones. Consideremos la gráfica en (Fig. 2).

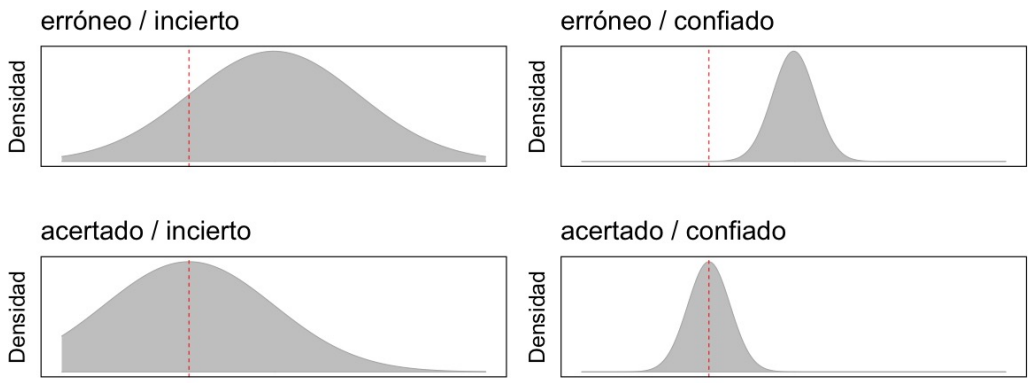


FIGURA 2. *Predicciones probabilísticas.*

En la Fig. 2 tenemos 4 modelos tratando de predecir una cantidad no observada (la línea punteada vertical). Claramente hay ciertas situaciones que son preferibles a otras. Por ejemplo, entre los dos modelos `confiados` preferimos el modelo `acertado`. Sin embargo, entre los dos modelos `erróneos` preferimos el modelo `incierto`. Entre los dos modelos `acertados` preferimos que el tiene mayor confianza en sus predicciones. Si realizamos las comparaciones veremos que preferiremos un modelo y con predicciones precisas. Mientras que un modelo erróneo e incierto sólo será preferido si se contrasta con un modelo con erróneo y confiado.

Lo que vemos es que para contrastar un modelo necesitamos una manera de poder comparar al mismo tiempo `asertividad` y `certidumbre`. Justo podríamos comparar en términos de las densidades. Y por cuestiones numéricas argumentar sobre comparar basados en la `log-densidad`.

Resulta que comparar con log-densidades es una regla que tiene muchas propiedades teóricas atractivas [4]. Lo que queremos es poder medir la log-densidad y compararla contra un oráculo que reporte las probabilidades verdaderas. La respuesta la encontramos en `teoría de la información` [5].

2.1. Información e incertidumbre

La pregunta que queremos resolver es: *¿Qué tanto se ha reducido mi incertidumbre cuando observo un resultado?*

La función que nos ayuda a medir la incertidumbre reflejada en un función de probabilidad es la `entropía` de dicha función de probabilidad. Supongamos que tenemos n posibles observaciones, cada una ocurriendo con probabilidad p_i entonces la entropía es

$$H(p) = -\mathbb{E}\log p = -\sum_{i=1}^{n} p_i \log p_i \,. \tag{1}$$

Con esa medida nos gustaría medir la incertidumbre adicional por utilizar una distribución distinta. Esto lo logramos con la `divergencia de Kullback-Leibler`. La cual definimos como

$$\mathsf{KL}(p\|q) = \sum_{i=1}^{n} p_i(\log p_i - \log q_i) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) \,. \tag{2}$$

Nota que la divergencia de Kullback-Leibler la podemos escribir como

$$\mathsf{KL}(p\|q) = H(p,q) - H(q)\,, \tag{3}$$

donde $H(p,q) = -\sum_{i=1}^{n} p_i \log q_i$ es la `entropía cruzada`.

Supongamos que tenemos un evento binario con probabilidades $p = \{0.3, 0.7\}$. Consideremos q una función que asigna las probabilidades de dicho evento binario. Esto puede ser desde una $q = \{0.01, 0.99\}$ hasta una $q = \{0.99, 0.01\}$.

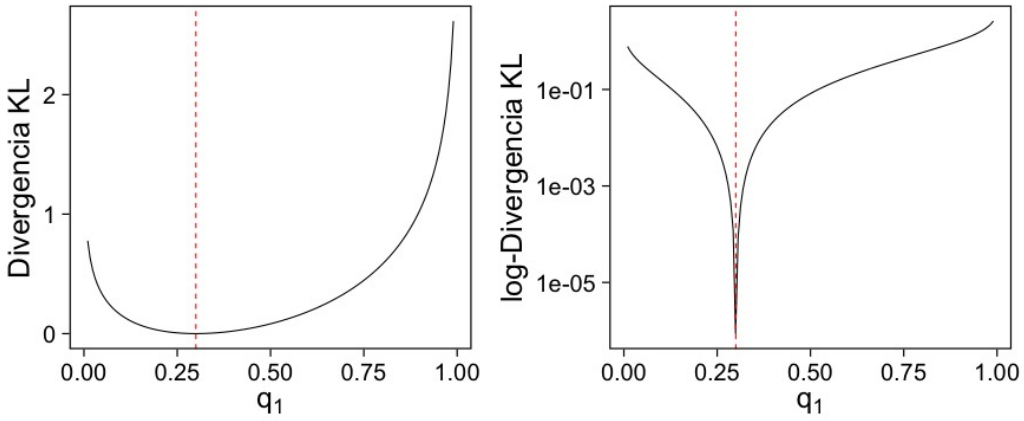


FIGURA 3. *Divergencia KL de usar q cuando los eventos ocurren con distribución p,* $\mathsf{KL}(p\|q)$.

En la práctica, por supuesto no conocemos p —no estaríamos haciendo inferencia— pero nos interesa poder contrastar dos modelos de probabilidad, q y r. Lo cual podemos realizar por medio de medir diferencias:

$$\mathsf{KL}(p\|q) - \mathsf{KL}(p\|r)\,. \tag{4}$$

¿Qué es lo que notas de la ecuación de arriba?

Calcular la diferencia elimina el término $H(p)$ y al final no lo necesitamos. El término que queda es el referente a los términos de `entropía cruzada` entre nuestros modelos de probabilidad y el mecanismo que genera los datos p.

2.2. En el contexto Bayesiano

Hemos establecido que podemos calcular la entropía cruzada de cada modelo para poder comparar entre alternativas. Para esto necesitamos calcular las log-densidades bajo nuestro modelo bayesiano. Esto es, necesitamos calcular `log-densidad predictiva posterior puntual` en $\tilde{y}_i$

$$\mathsf{lppd}(\tilde{y}_i) := \log \pi(\tilde{y}_i|\underline{y}_n) = \log \int \pi(\tilde{y}_i|\theta)\,\pi(\theta|\underline{y}_n)\,\mathrm{d}\theta\,. \tag{5}$$

Notemos que estamos promediando el proceso generador de datos (verosimilitud) con respecto a las posibles configuraciones que tienen sentido través de la distribución posterior.

Podemos ir mas allá y establecer el cálculo del `valor esperado de la log-densidad predictiva` en $\tilde{y}_i$, o mejor aún, en una colección de realizaciones

$$\mathsf{elppd} = \sum \int \pi(\tilde{y}_i) \log \pi(\tilde{y}_i|\underline{y}_n) \mathrm{d}\tilde{y}_i \,, \tag{6}$$

donde estamos utilizando nuestra distribución predictiva posterior para un conjunto de datos nuevo $\tilde{y}_i$, después de haber observado un conjunto de datos $\underline{y}_n$.

Nota que la expresión de elpd evalúa la capacidad predictiva del modelo en términos de la log-verosimilitud de manera puntual en cada una de nuevas muestras. El problema es nuestro desconocimiento de $\pi(\tilde{y}_i)$.

2.3. Consideraciones prácticas

En la expresión anterior estamos haciendo uso de una distribución para datos nuevos ($\pi(\tilde{y})$) la cual no conocemos. Así que lo que hacemos es calcular un resumen de la `log-densidad predictiva posterior puntual` evaluada en nuestros datos

$$\mathsf{lppd}(n) := \sum_{i=1}^{n} \mathsf{lppd}(y_i) = \sum_{i=1}^{n} \log \pi(y_i|\underline{y}_n) \,. \tag{7}$$

Para la cual podemos utilizar un estimador Monte Carlo

$$\widehat{\mathsf{lppd}}(n) = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} \pi(y_i|\theta^s) \right) \,, \tag{8}$$

donde $\theta^s \sim \pi(\theta|\underline{y}_n)$.

2.4. Precuaciones

El estimador construido arriba tiene el riesgo de dar valores pueden ser muy optimistas. ¿Por qué?

3. MÉTODOS DE COMPARACIÓN DE MODELOS

En cualquier tarea de modelado predictivo nos interesa poder evaluar la capacidad de generalización del modelo construido. Usualmente utilizaríamos un conjunto de datos distinto o un conjunto de datos que veremos en un futuro cercano para evaluar la capacidad predictiva. Pero **bajo el contexto Bayesiano** al momento de hacer inferencia sólo podemos considerar un conjunto de datos para el ajuste. Además, como hemos mencionado antes **no conocemos** el mecanismo de cómo de se generan los datos.

Los mecanismos usuales para medir la capacidad predictiva de un modelo son:

1. *Capacidad predictiva dentro de muestra.*
2. *Capacidad ajustada dentro de muestra.*
3. *Validación cruzada.*

3.1. Criterio de información de Akaike (AIC)

El criterio de información de Akaike es el método tradicional para evaluar la capacidad predictiva general del modelo sin tener en consideración un conjunto de datos adicional. La métrica penaliza por el número de parámetros ([1, 3],) a través de

$$\widehat{\mathsf{elpd}}_{\mathsf{AIC}} = \log \pi(\underline{y}_n|\hat{\theta}_{\mathsf{MLE}}) - k \,, \tag{9}$$

donde k es el número de parámetros del modelo.

Nota que en la literatura es usual encontrar la expresión

$$\mathsf{AIC} = -2 \cdot \widehat{\mathsf{elpd}}_{\mathsf{AIC}} = -2 \log \pi(\underline{y}_n | \hat{\theta}_{\mathsf{MLE}}) + 2k \,, \tag{10}$$

donde en lugar de tenerlo escrito en términos de la *log densidad predictiva* (tema del curso) está definido en términos de devianza (en [10] se argumenta por el factor de -2 para tener una distribución asintótica χ^2 para una diferencia de devianzas).

3.2. Criterio de información de Devianza (DIC)

El criterio de información de Devianza (DIC) incorpora dos cambios en el contexto bayesiano. Reemplaza el estimador de `MLE` por un estimador bayesiano y el término relacionado a los parámetros se cambia por un estimado utilizando los datos. La métrica de capacidad predictiva es

$$\widehat{\mathsf{elpd}}_{\mathsf{DIC}} = \log \pi(\underline{y}_n | \hat{\theta}_{\mathsf{Bayes}}) - p_{\mathsf{DIC}} \,, \tag{11}$$

donde $\hat{\theta}_{\mathsf{Bayes}}$ es la media posterior y p_{DIC} es el `número efectivo de parámetros`.

El número efectivo de parámetros se puede calcular por medio de dos expresiones:

$$p_{\mathsf{DIC}} = 2 \left(\log \pi(\underline{y}_n | \hat{\theta}_{\mathsf{Bayes}}) - \mathbb{E}_{\theta | \underline{y}_n} [\log \pi(\underline{y}_n | \theta)] \right) \tag{12}$$

ó

$$p_{\mathsf{DIC}} = 2 \mathbb{V}_{\theta | \underline{y}_n} (\log \pi(\underline{y}_n | \theta)) \,. \tag{13}$$

Ambas estimaciones dan el resultado correcto en el límite de un modelo con número de parámetros fijos una colección grande de datos.

3.3. Criterio de información Watanabe-Akaike (WAIC)

El criterio de Watanabe-Akaike (WAIC) utiliza la log-densidad predictiva posterior puntual (lppd) y utiliza una corrección por el número efectivo de parámetros

$$p_{\mathsf{WAIC}} = \sum_{i=1}^{n} \mathbb{V}_{\theta | \underline{y}_n} (\log \pi(y_i | \theta)) \,, \tag{14}$$

por lo que la métrica la calculamos por medio de

$$\widehat{\mathsf{elppd}}_{\mathsf{WAIC}} = \widehat{\mathsf{lppd}}(n) - p_{\mathsf{WAIC}} \,. \tag{15}$$

Nota que es una métrica que necesita la log-densidad predictiva posterior puntual en cada una de las observaciones. Por detrás esto supone cierta estructura de independencia condicional de los datos. Se puede calcular para datos con cierta estructura (temporal o geográfica) pero es no es posible interpretar el resultado.

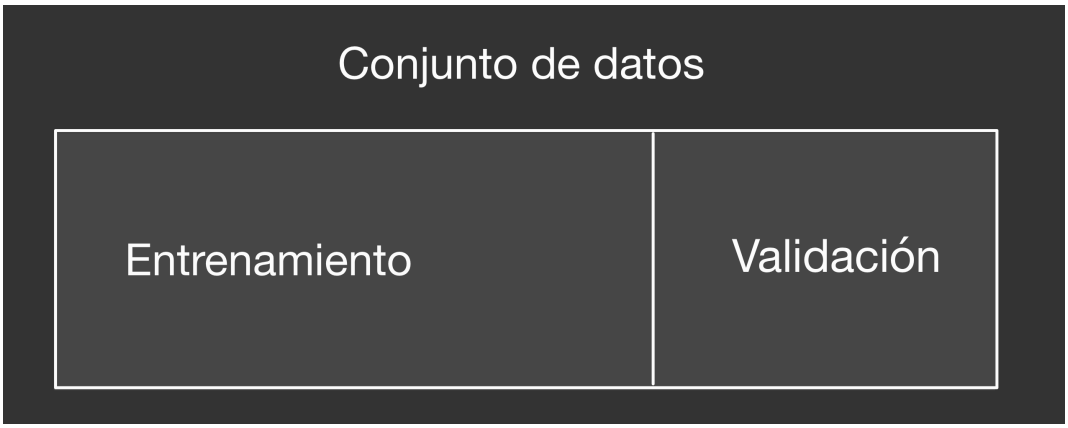


FIGURA 4. *Esquema de validación por separación de muestras.*

3.4. Validación cruzada

En modelado predictivo es usual partir los datos de tal manera que tengamos un conjunto para ajustar un modelo y un conjunto para estimar la capacidad predictiva de dicho modelo.

En la práctica no queremos dejar fuera los datos que tenemos para ajustar un modelo. Por lo tanto, lo que se usa es dividir el conjunto de datos en bloques. La idea es registrar el error de generalización (o alguna métrica adecuada de capacidad predictiva) cuando dejamos un bloque fuera del ajuste. Esto lo repetimos para cada bloque.

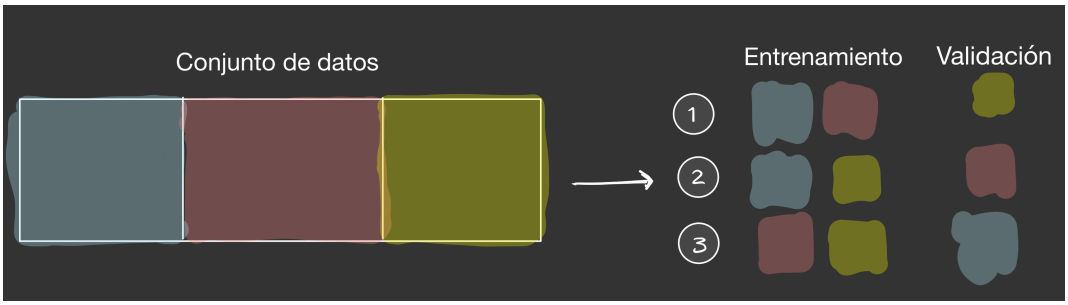


FIGURA 5. *Esquema validación cruzada con tres bloques.*

El caso extremo es considerar tantos bloques como observaciones tengamos (*leave-one-out cross validation*, `LOO-CV`). Aunque es un procedimiento costoso, existen diversas técnicas que permiten el cálculo del modelo completo y un ajuste por los `pesos por importancia` de cada una de las observaciones.

La capacidad predictiva con `LOO-CV` se calcula como

$$\widehat{\mathsf{lppd}}_{\mathsf{LOO}}(n) = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} \pi(y_i|\theta^s_{-i}) \right) , \tag{16}$$

donde $\theta^s_{-i} \sim \pi(\theta|y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$.

Muestreo por importancia nos permite calcular la capacidad predictiva utilizando pesos

$$w_s = \frac{1}{\pi(y_i|\theta^s)}, \qquad \theta \sim \pi(\theta^s|\underline{y}_n)\,. \tag{17}$$

para escribir

$$\widehat{\mathsf{lppd}}_{\mathsf{IS}}(n) = \sum_{i=1}^{n} \log \left(\sum_{s=1}^{S} \bar{w}_s \, \pi(y_i|\theta^s) \right), \qquad \bar{w}_s = \frac{w_s}{\sum_{k=1}^{S} w_k} \,. \tag{18}$$

Lo que puede suceder es que existan algunos `pesos mas grandes` que los demás y que dominen el cálculo de la ecuación anterior. Por lo tanto, la estrategia de [9] es suavizar los pesos mas grandes de acuerdo a una distribución Pareto generalizada:

$$\pi(r|u, \sigma, k) = \sigma^{-1}(1 + k(r-u)\sigma^{-1})^{-\frac{1}{k}-1} \,, \tag{19}$$

donde u es una cota inferior, σ un parámetro de escala (positivo), y k un parámetro de forma.

Con el método de suavizamiento podemos estimar los parámetros de la distribución Pareto (para observación). En particular, el parámetro k es el más informativo. Pues, nos da una indicación de que tan confiable es la aproximación.

La distribución Pareto tiene una varianza infinita si $k > 0.5$ que implica una distribución con colas pesadas. Como nos interesan los pesos y queremos suavizar los más grandes entonces buscamos que $k < 0.7$ (esto está bien fundamentado teorica y prácticamente, pueden consultar las referencias de [9]).

3.5. Ejemplo: modelo jerárquico

Regresaremos a nuestro ejemplo estrella del curso: los datos de la pruebas estandarizadas en las escuelas. Utilizaremos tres modelos posibles:

1. Modelo de parámetros independientes (*no pooling*).
2. Modelo de parámetros agrupados (*complete pooling*).
3. Modelo jerárquico.

Los datos que utilizaremos son los de [8].

```
## Caso: escuelas --------------------------------------
data ← tibble( id = factor(seq(1, 8)),
               y = c(28, 8, -3, 7, -1, 1, 18, 12),
               sigma = c(15, 10, 16, 11, 9, 11, 10, 18))

data.list ← c(data, J = 8)
```

Pondremos a prueba los tres modelos mencionados. Empezaremos con un modelo parámetros independientes. Esto es,

$$y_j \sim \mathsf{N}(\theta_j, \sigma_j) \,, \tag{20}$$

$$\theta_j \sim \mathsf{Constante} \,. \tag{21}$$

```
data {
  int<lower=0> J;
  real y[J];
  real<lower=0> sigma[J];
}
parameters {
  real theta[J];
}
model {
  y ~ normal(theta, sigma);
```

```
}
generated quantities {
  array[J] real log_lik;
  for (jj in 1:J){
    log_lik[jj] = normal_lpdf(y[jj] | theta[jj], sigma[jj]);
  }
}
```

Calcularemos las métricas de capacidad predictiva. Pero antes, tenemos que hacer un preprocesamiento. Necesitamos tener de nuestras muestras la evaluación de $\log \pi(y_j|\theta^s)$ y además la `eficiencia relativa del muestreador`.

```
library(loo)
posterior.indep ← modelo.indep$sample(data.list, refresh = 500)
stanfit ← rstan::read_stan_csv(posterior.indep$output_files())
log_lik ← extract_log_lik(stanfit, merge_chains = FALSE)
r_eff ← relative_eff(exp(log_lik), cores = 2)
```

Podemos calcular el `WAIC`:

```

Computed from 4000 by 8 log-likelihood matrix

          Estimate  SE
elpd_waic    -34.1 0.7
p_waic         4.0 0.2
waic          68.3 1.5

8 (100.0%) p_waic estimates greater than 0.4. We recommend trying loo instead.
Warning message:

8 (100.0%) p_waic estimates greater than 0.4. We recommend trying loo instead.
```

Vehtari y coautores –puedes ver las referencias sugeridas en el FAQ de `Stan`– recomiendan utilizar estimadores de `LOO-CV` pues junto con el procedimiento de suavizamiento Pareto otorga mejores diagnósticos de la estimación:

```

Computed from 4000 by 8 log-likelihood matrix

         Estimate  SE
elpd_loo    -36.4 0.8
p_loo         6.3 0.4
looic        72.8 1.7
------
Monte Carlo SE of elpd_loo is NA.

Pareto k diagnostic values:
                         Count Pct.    Min. n_eff
(-Inf, 0.5]   (good)     0       0.0%   <

 (0.5, 0.7]   (ok)       0       0.0%   <

   (0.7, 1]   (bad)      7      87.5%   18
   (1, Inf)   (very bad) 1      12.5%   23
See help('pareto-k-diagnostic') for details.
Warning message:
```

```
Some Pareto k diagnostic values are too high. See help('pareto-k-diagnostic')
    for details.
```

Ahora probemos un modelo completamente agrupado

$$y_j \sim \mathsf{N}(\theta, \sigma_j)\,, \tag{22}$$

$$\theta \sim \mathsf{N}(\mu, \tau)\,. \tag{23}$$

```
data {
  int<lower=0> J;
  real y[J];
  real<lower=0> sigma[J];
}
parameters {
  real mu;
  real<lower=0> tau;
  real theta_tilde;
}
transformed parameters {
  real theta = mu + tau * theta_tilde;
}
model {
  mu ~ normal(0, 5);
  tau ~ cauchy(0, 5);
  theta_tilde ~ normal(0, 1);
  y ~ normal(theta, sigma);
}
generated quantities {
  array[J] real log_lik;
  for (jj in 1:J){
    log_lik[jj] = normal_lpdf(y[jj] | theta, sigma[jj]);
  }
}
```

Y también pondremos a prueba nuestro modelo jerárquico estudiado antes.

```
data {
  int<lower=0> J;
  real y[J];
  real<lower=0> sigma[J];
}
parameters {
  real mu;
  real<lower=0> tau;
  real theta_tilde[J];
}
transformed parameters {
  real theta[J];
  for (j in 1:J)
    theta[j] = mu + tau * theta_tilde[j];
}
model {
  mu ~ normal(0, 5);
  tau ~ cauchy(0, 5);
  theta_tilde ~ normal(0, 1);
  y ~ normal(theta, sigma);
}
```

```
generated quantities {
  array[J] real log_lik;
  for (jj in 1:J){
    log_lik[jj] = normal_lpdf(y[jj] | theta[jj], sigma[jj]);
  }
}
```

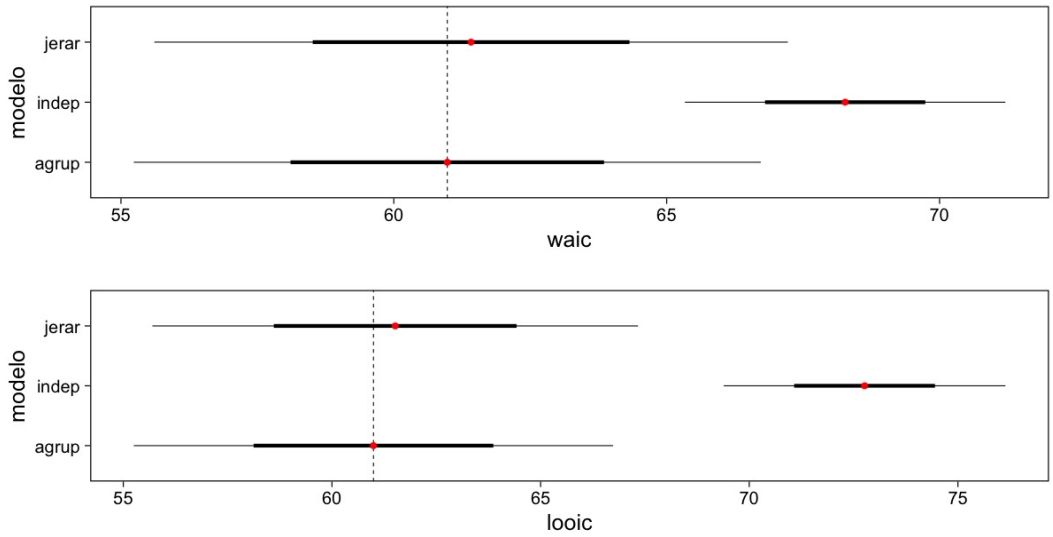
Podemos comparar de manera puntual cada modelo por medio de `WAIC`

```
       elpd_diff se_diff elpd_waic se_elpd_waic p_waic se_p_waic waic  se_waic
agrup    0.0       0.0    -30.5       1.4          0.5    0.2       61.0   2.9
jerar   -0.2       0.2    -30.7       1.5          0.9    0.3       61.4   2.9
indep   -3.6       1.3    -34.1       0.7          4.0    0.2       68.3   1.5
```

O podemos comparar por medio de `LOO-PSIS`

```
       elpd_diff se_diff elpd_loo se_elpd_loo p_loo se_p_loo looic se_looic
agrup    0.0       0.0    -30.5      1.4         0.5    0.2      61.0   2.9
jerar   -0.3       0.2    -30.8      1.5         0.9    0.3      61.5   2.9
indep   -5.9       1.4    -36.4      0.8         6.3    0.4      72.8   1.7
```

Los resultados son muy similares bajo ambos métodos. Sin embargo, `LOO-PSIS` nos provee de mejores diagnósticos en el cómputo de la capacidad predictiva del modelo



Nota que no hemos calculado el `AIC` para este modelo, ¿por qué?

REFERENCIAS

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, Hungary, 1973. Akademiai Kiado. 4

[2] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*, volume 2. CRC press Boca Raton, FL, 2014. 1

[3] A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, nov 2014. ISSN 0960-3174, 1573-1375. . 1, 4

[4] T. Gneiting and A. E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, mar 2007. ISSN 0162-1459, 1537-274X. . 2

[5] D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003. 2

[6] O. A. Martin, R. Kumar, and J. Lao. *Bayesian Modeling and Computation in Python*. Chapman and Hall/CRC, Boca Raton, First edition, 2021. 1

[7] R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Texts in Statistical Science. Taylor and Francis, CRC Press, Boca Raton, Second edition, 2020. ISBN 978-0-367-13991-9. 1

[8] D. B. Rubin. Estimation in Parallel Randomized Experiments. *Journal of Educational Statistics*, 6(4): 377–401, 1981. ISSN 0362-9791. . 7

[9] A. Vehtari, D. Simpson, A. Gelman, Y. Yao, and J. Gabry. Pareto Smoothed Importance Sampling. *arXiv:1507.02646 [stat]*, feb 2021. 7

[10] S. S. Wilks. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, mar 1938. ISSN 0003-4851, 2168-8990. . 5