UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

# Evaluación de un programa orientado a promover el desarrollo de competencias ciudadanas y emocionales en instituciones educativas.

**Fernando López-Torrijos Flórez**

Universidad Santo Tomás
Facultad de Estadística
División de Ciencias Económicas y Administrativas
Bogotá, D.C., Colombia
2019

# Evaluación de un programa orientado a promover el desarrollo de competencias ciudadanas y emocionales en instituciones educativas.

**Fernando López-Torrijos Flórez**

Trabajo de grado presentado como requisito parcial para optar al título de:
**Magister en Estadística Aplicada**

Director:
Juan Camilo Sosa Martínez (Ph.D.)

Universidad Santo Tomás
Facultad de Estadística
División de Ciencias Económicas y Administrativas
Bogotá, D.C., Colombia
2019

# Resumen

Evaluar las intervenciones en el sector social es una necesidad sentida por parte de los financiadores para verificar la eficacia de sus aportes, por parte de la población intervenida para verificar que su esfuerzo es compensado y por parte de las entidades coordinadoras y ejecutoras para la divulgación y promoción de sus servicios. La evaluación cuantitativa utiliza usualmente técnicas de la estadística frecuentista: diferencia de medias para comparar dos poblaciones, ANOVA cuando se comparan las medias de más de dos poblaciones y modelos lineales cuando hay otras variables en juego.

La presente evaluación se realiza aplicando modelos lineales mixtos, en específico modelos lineales jerárquicos, en tres niveles, bajo la perspectiva Bayesiana, primero sin cofactores y luego adicionándolos. La aplicación práctica se realiza alrededor de la evaluación de un programa orientado a promover el desarrollo de competencias ciudadanas y emocionales en instituciones educativas oficiales, en el nivel de educación básica secundaria, por medio del área de Educación Física, Recreación y Deporte.

Palabras clave: Estadística Bayesiana, evaluación, modelo jerárquico, modelo lineal mixto, educación secundaria, competencias ciudadanas, competencias emocionales, evaluación de programas sociales, evaluación de políticas públicas.

# Abstract

Evaluating interventions in the social sector is a felt need by the funders to verify the effectiveness of their contributions, by the population involved to verify that their effort is compensated and by the coordinating and executing entities for the dissemination and promotion of their services. The quantitative evaluation usually uses techniques of frequentist statistics: difference of means to compare two populations, ANOVA when comparing the means of more than two populations, and linear models when there are other variables at play.

This evaluation is carried out by applying mixed linear models, in specific hierarchical linear models, on three levels, under the Bayesian perspective, first without cofactors and then adding them. The practical application is carried out around a program assesment aimed at promoting the development of citizen and socio-emotional skills in official secondary educational institutions through the area of Physical Education, Recreation and Sports.



Keywords: Bayesian statistics, evaluation, hierarchical model, mixed linear model, secondary education, citizen competencies, emotional competencies, social programs assesment, public policies assesment.

# 1. Introducción

Hay muy diversas propuestas acerca de cómo promover la formación ciudadana y socioemocional en las escuelas, reconociendo que muchos otros espacios de socialización también tienen que cumplir un papel importante, pero se menciona a la escuela como espacio fundamental de formación: "Todas las interacciones entre estudiantes o entre adultos y estudiantes, la construcción de normas que regulan esas interacciones, las decisiones que se toman, los conflictos y problemas que surgen, casi todo lo que ocurre de manera cotidiana en la escuela puede ser tomado como oportunidad para la formación ciudadana" y "... la formación ciudadana puede estar perfectamente relacionada con la formación académica que ocurre permanentemente en la escuela"[1].

La Organización de Estados Americanos (OEA) proclama que "en el proceso de desarrollo de la educación ciudadana, podemos observar que la idea de que el deporte representa una valiosa herramienta de educación democrática, ha tomado fuerza tanto en el continente americano como en el resto del mundo. Esta tendencia descansa sobre la premisa que el deporte constituye, más allá de un ámbito de entrenamiento físico, un espacio idílico de interacción para que jóvenes de diferentes orígenes sociales, religiosos o étnicos se conozcan y aprendan a convivir. En pocas palabras, «si pueden jugar juntos, pueden vivir juntos». En consecuencia, en los últimos años se viene observando una multiplicación de Organizaciones No Gubernamentales que utilizan la práctica deportiva como vehículo para educar en valores como el compromiso comunitario (Goals Haiti), la inclusión social (United Soccer Club) o el liderazgo juvenil (Fundación Tiempo de Juego)"[17].

Las intervenciones evaluadas conjugan las iniciativas que utilizan el deporte como espacio para el desarrollo de competencias personales y sociales con la oportunidad que da el área de educación física, recreación y deporte de los establecimientos educativos para desarrollarlas. Las secretarias de Educación, en alianza con Fundaciones de origen empresarial, han implementado programas en el nivel de básica secundaria en diversas instituciones educativas con el objeto de que las canchas se conviertan en un escenario de aprendizajes significativos que aporten a la convivencia escolar y a la construcción de una sociedad en paz.

Se enfocan en el apoyo de actividades que contribuyan a la adquisición de habilidades sociales, recursos asertivos para la resolución de conflictos en la escuela y el fomento en general de otros valores que se desenvuelven en lo que se denomina *desarrollo juvenil positivo*, el cual se podría definir "como la participación en comportamientos prosociales y la prevención de comportamientos que comprometan la salud y futuras conductas peligrosas" [12][1]

La intervención proporciona a los docentes de educación física, recreación y deporte de dichas instituciones una formación teórico-práctica certificada para conocer la metodología de cada programa y cómo introducirla en los ejercicios y actividades de las sesiones con sus alumnos. Las metodologías suelen explicitar las competencias ciudadanas y habilidades para la vida que pretende fomentar cada sesión, de tal manera que sirvan de apoyo para preparar la sesión, para desarrollarla y para su evaluación posterior. Adicionalmente, el proyecto puede proveer entrenadores especializados que acompañan algunas sesiones, con el objeto de afianzarles el manejo de la metodología a los docentes.

Los programas que se desarrollan en entidades territoriales a lo largo y ancho del país tienen financiadores, presupuesto gubernamental y aportes de entidades sin ánimo de lucro, que necesitan conocer si los recursos están logrando los objetivos para los cuales fueron destinados y en qué medida lo logran. Se trata de rendir cuentas del programa a estos financiadores, pero también a

[1] Holt (2008) citando 'Promoting healthy adolescents: synthesis of youth development program evaluations', Journal of Research on Adolescence, 8: 423–59. Roth, J., Brooks-Gunn, J., Murray, L., and Foster, W. (1998), pag. 2

las instituciones educativas que implementan el programa, ya que invierten tiempo y energía en el desarrollo del mismo y les es importante también rendir cuentas a los padres de familia y a los superiores inmediatos. Las entidades coordinadoras y ejecutoras desean divulgar y promocionar sus servicios. También se trata de tener elementos de juicio que permitan determinar, de entre un abanico de alternativas, cuáles rinden mayores beneficios, entendidos éstos de manera amplia: financieros, sociales o formativos. Las mejores alternativas son aquellas que se divulgan como mejores prácticas a nivel nacional, continental o mundial y las que planes, programas y proyectos futuros replican buscando su adaptación y mejora. Pero nada de esto es posible sin la realización de evaluaciones técnicas. La evaluación de los programas sociales y de las políticas públicas es un aspecto fundamental para la mejora de las intervenciones llevadas a cabo sobre una comunidad. Evaluar permite que los gobiernos, organizaciones no gubernamentales y organismos internacionales focalicen adecuadamente sus recursos y elaboren programas que generen bienestar en las poblaciones atendidas por los programas sociales.

Cada programa desarrolla su sistema de evaluación cualitativa y cuantitativa, si bien no hay una cultura de evaluación cimentada alrededor de este tipo de programas[15]. Las competencias son factores latentes de los cuales el estudiante no necesariamente es consciente. Para la parte cuantitativa se han desarrollado instrumentos de autoconcepto o que dan cuenta de la medida en que han incorporado las competencias en su quehacer y ser cotidiano. Los estudiantes los diligencian al inicio del programa y al finalizar la intervención. Se asume que responden con honestidad, puesto que no hay respuestas correctas ni incorrectas. El propósito de las herramientas es indagar el nivel de desarrollo de las competencias socioemocionales y ciudadanas mencionadas en el periodo de tiempo desde que se inicia hasta que se acaba el proyecto. Los estudiantes están identificados en cada instrumento, de tal modo que se tienen mediciones pareadas. Las respuestas permiten construir y asignar un índice individual para las competencias ciudadanas o habilidades para la vida que pretende desarrollar la intervención. La comparación del índice entre cada par de medidas pareadas permite calcular un valor de diferencia entre un momento final, una vez terminada la intervención y un momento inicial, previo a la intervención, por alumno.

No se trata de una evaluación de impacto ya que la intervención no cumple con los elementos necesarios para poder determinar causalidad, es decir, atribuir el cambio específicamente a la intervención, ya que, por ejemplo, no se seleccionaron ni midieron establecimientos educativos de control. No hay contrafactuales. Tampoco hay una selección aleatorizada de las instituciones educativas. Éstas se seleccionan en un acuerdo entre la Secretaría de Educación, cada rector y los docentes del área de educación física, recreación y deporte de tal forma que en la decisión median decisiones políticas en el buen sentido de la palabra.

La evaluación de programas basada en evidencia hace uso de metodologías cuantitativas con enfoque frecuentista y bayesiana que permiten cuantificar los impactos del programa a través de metodologías como diferencias de medias y proporciones, diferencias en diferencias, regresión discontinua, entre otras. En esta evaluación se cuantifica la diferencia de medias entre el momento en que finalizó el programa y aquel en que inició, a nivel global y diferenciando por sexo o por grado, por medio del uso de modelos lineales mixtos, en particular se desarrollará un modelo jerárquico de tres niveles bajo la perspectiva Bayesiana.

Bajo la perspectiva utilizada no se puede aducir causalidad de ningún modo. El programa contiene una mezcla los factores intrínsecos, tales como la maduración y la motivación del individuo, y extrínsecos, tales como los estímulos del medio ambiente. La intervención es uno de esos estímulos externos. Todos los elementos suman o restan al resultado y no es posible separarlos.

Vale la pena puntualizar que se utiliza la palabra *evaluación* desde el punto de vista de la economía, no de la pedagogía, ya que para ésta los datos cuantitativos expresan una *medición*, mientras la *evaluación* expresa la interpretación o la emisión de juicios de valor sobre los datos aportados o resultantes de la medición [10] (Gutierrez, pag 12).

Se cuenta con información de varias aulas por cada establecimiento educativo. Y para cada individuo, se conoce el sexo, grado y aula a la que pertenece.

Las intervenciones sociales de este tipo que se realizan por fuera del sector educativo suelen discriminar los grupos según sexo. En cambio, en el contexto escolar se aplica en contextos mixtos, lo cual hace interesante el análisis por sexo. Del mismo modo, los estudiantes son bastante homogéneos en edad, permitiendo estudiar a qué edades es más efectiva la intervención. Pero el sector educativo no controla las edades de los jóvenes que participan en la intervención, sino los grados en los que se aplica, de tal modo que interesa un análisis por grado.

Los trabajos cuantitativos publicados en relación al nivel de la educación básica y media se concentran fundamentalmente en estudios del valor agregado o al modelamiento de los factores asociables a la calidad a partir de resultados en las pruebas estandarizadas en las áreas de lenguaje, matemáticas, ciencias sociales y ciencias naturales que financian o realizan entidades del nivel nacional, como el Instituto para la Evaluación de la Educación (ICFES) en Colombia, o estudios internacionales como PISA - un programa de la Organización para el Desarrollo y la Cooperación Económica (OECD) que se efectúa en múltiples países, y que evalúa conocimientos y habilidades relacionados con los dominios de comprensión lectora, matemática y científica dirigidas a jóvenes de 15 años que estén cursando al menos grado 7° - o SERCE - un proyecto del Laboratorio Latinoamericano de la Evaluación de la Calidad de la Educación (LLECE) de la OREALC/UNESCO, que evalúa competencias básicas y habilidades en las áreas de lectura, matemática y ciencias naturales dirigido a estudiantes de grados 3° y 6° de países latinoamericanos -. Pero hay muy poca literatura asociada a pruebas de desempeño en habilidades para la vida o competencias ciudadanas y competencias emocionales. Gutierrez menciona que el Plan Nacional de Educación colombiano del 2008 establece que la evaluación integral debe considerar además de la adquisición de aprendizajes en las dimensiones del pensamiento, la lectoescritura, las ciencias naturales y las matemáticas, asuntos relacionados con el desarrollo socio-afectivo, la corporalidad y la formación en valores; sin embargo "no hay referentes claros que permitan incorporar éstos factores de evaluación en la vida escolar y tenerlos en cuenta tanto para el mejoramiento integral de la calidad como de la promoción (Plan Decenal, 2008, pag 7); en este sentido, es fundamental preguntarse si la Educación Física constituye una estrategia de evaluación en estos campos". (pag 10)[10].

Casey & Goodyear revisaron 27 estudios relacionados con la formación de habilidades para la vida a través del deporte por medio de la metodología de aprendizaje colaborativo. Reportan un equilibrio entre los procedimientos de recopilación y análisis de datos cualitativos y cuantitativos (14 estudios utilizaron métodos cualitativos, 11 diseños cuantitativos y 2 métodos mixtos) y que sólo los estudios cuantitativos compararon el aprendizaje de los estudiantes con un grupo de control. Una de sus conclusiones finales es que si bien el aprendizaje colaborativo ha tenido un comienzo prometedor en el área de la educación física, todavía tiene mucho que demostrar[2].

Este trabajo está estructurado como sigue: en la Sección 2 se exponen las motivaciones asociadas al planteamiento del trabajo y las preguntas a las que se espera dar respuesta; en la Sección 3 se realiza un breve recuento de la literatura asociada a la técnica de modelado jerárquico; en la Sección 4 se proponen tres modelos, dos de ellos para dar respuesta a las preguntas y se discute la forma en que se validarán, donde además se discute la validez del supuesto de intercambiabilidad; en la Sección 5

se consideran los aspectos relacionados con la implementación y validación de los modelos junto con la elicitación de los hiperparámetros; en la Sección 6 se presentan los resultados correspondientes a tres índices sociogrupales que se toman como ejemplo para la discusión; y finalmente, en la Sección 7 se presentan conclusiones y algunas alternativas de investigación futura.

## 2. Planteamiento del problema

Las entidades gubernamentales, principalmente secretarías de las entidades territoriales o el ministerio nacional, están interesadas en evaluar si las propuestas de intervención llevadas a cabo por organizaciones no gubernamentales logran los objetivos trazados, por ende, solicitan estimar si se obtiene suficiente evidencia para afirmar que ha habido un cambio, ya sea de manera global, o por establecimiento educativo, y en qué dirección. Cada operador propone la metodología de estimación. Sólo les es exigido que se recopile información de un número representativo de alumnos a nivel global y que se genere un reporte de todos y cada una de las instituciones educativas. Algunos operadores aplican sus instrumentos de medición al número de alumnos acordado previamente con el financiador, a quienes seleccionan aleatoriamente. Las intervenciones suelen involucrar, en Colombia, entre 10 y 20 instituciones educativas.

Desde el punto de vista de la administración educativa, para determinar si una intervención propuesta por una organización no gubernamental vale la pena escalarla en forma de un programa educativo financiado por los recursos públicos, necesita determinar si la intervención es efectiva independientemente del contexto de las instituciones educativas, por cuanto una urbe metropolitana maneja gran diversidad de entornos y problemáticas.

La aproximación metodológica usual tiene problemas:

1. Se analiza el cambio en los estudiantes para concluir sobre los cambios en la institución. Realizar esto no es correcto ya que puede llevar a concluir que hay diferencia siendo que no la hay. El punto es que las pruebas de hipótesis estadísticas descansan en el supuesto de independencia de las observaciones y este supuesto es violado en estructuras escolares. Es razonable que los individuos de un mismo establecimiento educativo tengan más en común que individuos de diferentes establecimientos educativos, por razón del contexto social y geográfico o la creación de una cultura y clima institucional particular ocasionando que haya mayor heterogeneidad entre establecimientos educativos que intra establecimientos educativos. Los errores estándar estimados de las pruebas de hipótesis estadísticas tradicionales serán bastante reducidos, y esto conducirá a que la mayoría de los resultados sean espúreamente significativos[18], o en otras palabras, a sobre-reaccionar y decir que algo es efectivo siendo que no lo es.

2. Las personas que reciben el resultado pertenecen a las ciencias sociales y no están muy familiarizadas con los métodos cuantitativos. Cuando se realiza la evaluación bajo la perspectiva de la estadística frecuentista el mensaje transmitido por el evaluador y el mensaje entendido por los receptores no suele ser el mismo. Por ejemplo, la hipótesis nula de que no hay diferencia entre un par de mediciones implica que *con cierto nivel de confianza permite rechazar la hipótesis nula y por ende su valor esperado es tanto*. A veces se expone el intervalo de confianza y se les explica que se debe entender como *la región en donde existe un cierto nivel de confianza de que esté comprendido el valor esperado y que permite rechazar la hipótesis nula*. Los receptores a menudo interpretan erróneamente el valor p como *lo mismo que la probabilidad de la hipótesis alterna dada la evidencia* o que con el nivel de confianza expuesto *los datos no se han producido por causas del azar*, o que se les ha demostrado *el tamaño del efecto o la importancia del resultado*[20][5].

3. Aún bajo una adecuada interpretación de los resultados y una intención de aplicar un modelamiento adecuado a la estructura jerárquica del sector y a la aplicación de la intervención a grupos, no individuos, este tipo de intervenciones no debería ser evaluado por medio de métodos frecuentistas debido a que las intervenciones piloto cuentan usualmente con pocas

instituciones educativas y/o aulas. McCoach[14] hace referencia a simulaciones de modelamientos jerárquicos mediante modelos lineales mixtos en donde se requieren al menos 100 grupos (aulas) para una estimación razonable de los errores estándar del nivel del agrupamiento, de otro modo serían subestimados, y al menos 30 grupos (aulas) para producir estimadores insesgados de los componentes de varianza de dicho nivel, ya que de otro modo serían sobreestimados. No hay una compilación de las pruebas piloto que se han realizado en Colombia de tal manera que se pueda establecer una estadística de la cantidad de establecimientos educativos y aulas que suelen utilizarse para la exploración de la efectividad de las propuestas de intervención. El autor de este trabajo ha sido testigo de una docena de intervenciones, las cuales suelen abarcar de 10 a 15 instituciones educativas, y dentro de éstas no se implementa la intervención de manera generalizada en todas las aulas de los grados a las que está dirigida la intervención, sino a un pequeño grupo que puede oscilar entre 4 y 10 aulas por establecimiento educativo. A mayor número de establecimientos educativos, menor número de aulas por establecimiento. La razón es financiera, ya que no se cuenta con suficiente presupuesto para financiar una intervención de mayor tamaño.

La estadística Bayesiana es una alternativa de solución a la situación problemática ya que una vez realizado el modelamiento, permite responder las preguntas que realizan las partes interesadas en las intervenciones educativas: modela la estructura multinivel del sector, evitando resultados espúreos, el mensaje que emite tiene mayor posibilidad de que coincida con el que entienden los receptores de los resultados y logra evaluar intervenciones con un bajo número de aulas y/o establecimientos educativos.

## 2.1 Preguntas del estudio

De manera específica, la aplicación práctica del método Bayesiano busca:

1. Calcular cuál es la probabilidad de que el valor de la diferencia entre la medición una vez terminada la intervención y la medición previa sean mayores a cero.
2. Realizar una comparación entre los establecimientos educativos que participaron en el programa con el objetivo de determinar en qué medida varían los índices diferencia. Lo que interesa es determinar si la intervención se ve afectada por contextos individuales o logra efectos uniformes en todas las instituciones educativas, lo cual es preferible si se planteara una implementación que involucrara mayor cantidad de establecimientos.
3. Siendo que hay deportes que en el contexto de cada región o país son practicados mayoritariamente por un sexo, se desea calcular la diferencia según sexo para determinar si hay alguna brecha de género.
4. El programa se aplica a estudiantes de varios grados. La edad por grado está especificada por ley y se cumple en mayor o menor medida en cada establecimiento educativo. La niñez y adolescencia es una época de rápido cambio en los sujetos. Se desea calcular la diferencia por establecimiento educativo según grado, determinando si la intervención obtiene resultados más convenientes en unos grados que en otros.

## 2.2 Contribución del trabajo

Los métodos usualmente utilizados para la evaluación de este tipo de intervenciones puntuales son técnicas estadísticas frecuentistas que se enseñan comúnmente en un primer o segundo curso de estadística introductoria e intermedia. Se basan en la teoría normal y en el modelo lineal general,

un marco que incluye las pruebas t, la regresión lineal y el ANOVA. El paradigma del modelo lineal general asume independencia entre observaciones. Cuando se viola esta suposición, como ocurre con los datos en jerarquía del sector educativo o de mediciones repetidas, se necesitan técnicas estadísticas más avanzadas para tener en cuenta las dependencias de datos que surgen. Técnicas avanzadas como modelos mixtos y marginales o el modelado de ecuaciones estructurales rara vez se aplican. En la literatura hispanoamericana, tal vez sea en la Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en la Educación donde colaboran investigadores en educación que publican documentos utilizando técnicas modernas. Autores como Rubén Cervini, Emilio Blanco o Francisco Javier Murillo, entre una docena, son los que utilizan las técnicas multinivel bajo modelos de regresión lineal.

Su poca utilización puede deberse a que la capacitación en estos métodos solo está disponible en los programas académicos en los últimos años. Muchas de las técnicas estadísticas avanzadas que rara vez se observan en los estudios son métodos que no estaban disponibles en el software estadístico más accesible hacía diez o veinte años. Por ejemplo, los investigadores experimentados pueden no haber estado expuestos a técnicas de modelado estadístico modernizadas que ahora están disponibles y son adecuadas para analizar datos dependientes o de múltiples niveles. También el software estadístico y la potencia informática actual es la que democratiza el acceso y el uso de estos métodos estadísticos. Los métodos Bayesianos son otro tipo de técnicas que aprovechan la capacidad de computo actual y la disponibilidad de software para implementarlo. Realizar un ejercicio de modelado multinivel bajo el paradigma Bayesiano es una contribución a la difusión de las nuevas posibilidades.

Adicionalmente, en los casos en que se modela multinivel, lo usual es representarlo con dos niveles. Este trabajo se realiza modelando tres niveles, permitiendo realizar análisis a nivel aula y a nivel institución educativa, no teniendo que elegir uno de los dos.

Este estudio no hace uso de pruebas estandarizadas relacionadas con habilidades para la vida o competencias ciudadanas y/o emocionales de uso internacional, pero genera resultados cuantitativos que pueden ser útiles como referencia para evaluaciones futuras dada la poca disponibilidad de estudios documentados de este tipo de intervenciones.

Desde el punto de vista estadístico, el trabajo desarrolla la formulación explícita del modelo jerárquico de tres niveles, incluyendo expresiones completas para la distribución posterior, las distribuciones condicionales completas, y el algoritmo de Metropolis cuando no es cerrado. También realiza el desarrollo completo de los algoritmos por medio de MCMC para explorar una distribución posterior altamente dimensional sin utilizar programas precursores (como JAGS o BUGS) ni librerías, excepto para funciones anexas, tales como el cálculo de la autocorrelación de las cadenas. Por tanto, se tiene la materia prima para desarrollar un paquete enfocado al modelado jerárquico de tres niveles ya que se desarrolló todo el código parametrizado. Además, realiza la evaluación de los muestreadores por medio de simulación y la evaluación de los modelos mediante validación cruzada y criterios de información.

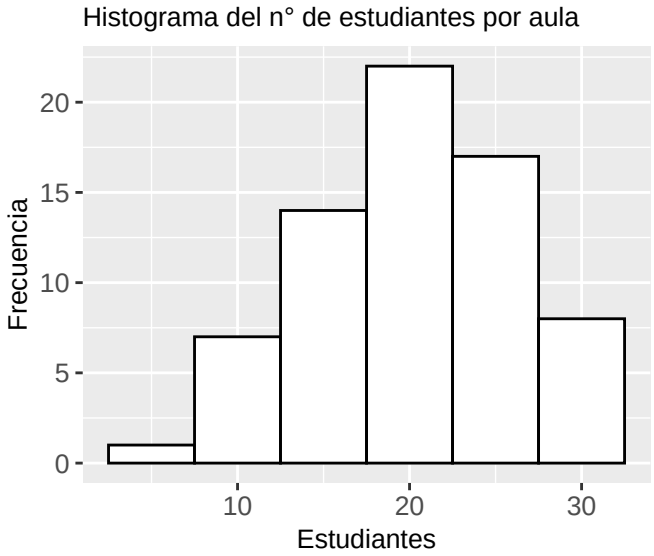
## 2.3 Información de contexto

La aplicación práctica del modelamiento se realiza sobre tres indices medidos en el año 2018 para veintiocho instituciones educativas oficiales de la ciudad de Bogotá. En cada una participaron entre 1 y 5 aulas, la mayoría con 2 ó 3 aulas, para un total de sesenta y nueve aulas, y más de mil trescientos estudiantes.

El número de estudiantes evaluados por aula varía entre 7 y 32 estudiantes.

Cuadro 1: Tabla de frecuencia de número de establecimientos educativos según número de aulas

| N° de aulas | N° de establecimientos educativos |
|---|---|
| 1 | 2 |
| 2 | 14 |
| 3 | 10 |
| 4 | 1 |
| 5 | 1 |




Los indices analizados son *Liderazgo efectivo*, *Habilidades comunicativas* y *Respeto hacia los demás y hacia sí mismo*. Las tres son habilidades socio-grupales.

Liderazgo efectivo entendido como la capacidad para ejercer liderazgo en la ejecución de una tarea.

Habilidades comunicativas definido como logro de una comunicación efectiva con pares y docentes, en contextos interpersonales y en grupo.

Respeto hacia los demás y hacia sí mismo en términos de que el joven muestra claridad acerca del comportamiento adecuado frente a los límites ligados al espacio personal, cuerpo y tiempo propios y de los demás. Presenta capacidad para respetarlos y hacerlos respetar de forma adecuada.

Los índices son una variable cuantitativa continua medida en escala de intervalo. Es una entrada que recibe el modelo. Cómo llegar desde preguntas cualitativas medidas en escala ordinal a una variable latente cuantitativa continua, el índice, está fuera del alcance de este trabajo.

Se tiene la medición pareada por estudiante de los índices en un momento inicial, antes de iniciar la intervención, y un momento final, al terminar ésta. Se toma como índice para el trabajo la diferencia entre este par de mediciones. Entiéndase por tanto que se utiliza un *indice diferencia*. A partir de la estadística descriptiva de los tres índices diferencia se observa que la media es muy cercana a cero y la desviación estándar a uno. Así mismo, la distribución de los índices diferencia se puede considerar simétrica y tendiente a mesocúrtica[2]. En consecuencia se considera que la distribución

[2]Ligeramente leptocúrtica

Cuadro 2: Estadística descriptiva de los índices de competencias ciudadanas

| | Liderazgo efectivo | Habilidades comunicativas | Respeto por los demás y por sí mismo |
|---|---|---|---|
| Mínimo | -3.600 | -4.333 | -4.000 |
| q1 | -0.400 | -0.500 | -0.600 |
| Mediana | 0.000 | 0.000 | 0.000 |
| Media | 0.097 | 0.061 | -0.064 |
| q3 | 0.600 | 0.667 | 0.600 |
| Máximo | 4.200 | 4.000 | 3.000 |
| Desv. Est | 1.030 | 0.972 | 0.889 |
| Sesgo | 0.195 | -0.037 | -0.041 |
| Curtosis | 1.169 | 1.175 | 0.836 |

teórica subyacente es gaussiana. El cuadro 2 proporciona la estadística descriptiva.

## 3. Revisión de literatura

Snijders y Bosker mencionan que en la forma actual el análisis multinivel tiene dos fuentes: el *análisis contextual* desarrollado por las ciencias sociales y el del modelamiento de *efectos mixtos*. Dentro de la primera fuente menciona que Robinbson (1950) discute la falacia ecológica en la que se confunde entre los efectos agregados y los efectos individuales, y que Davis et al. presenta la distinción entre varianza intra-grupo y la varianza inter-grupo en las regresiones, o el paper escrito por Burstein et al. (1978) que trata las pendientes y los interceptos de las regresiones en un nivel como resultados de un nivel superior. Para la segunda fuente considera importante el capítulo 2 de Searle et al. (1992) quienes exponen una repaso histórico del tema. Considera que las bases del análisis multinivel se establecieron en 1986, pero que a partir del año 2000 es que empezaron a florecer textos acerca de dicho tema. Como áreas de aplicación menciona las ciencias biomédicas usando datos longitudinales y la economía con datos panel. Dicen que se acude a este tipo de modelos cuando se está interesado en proposiciones acerca de variables que están conectadas a través de diferentes niveles[18].

Gelman et al. mencionan que el modelamiento lineal jerárquico ha ganado en popularidad últimamente (lo escriben en 2014), especialmente en las ciencias sociales, y que a menudo es denominado modelamiento multinivel[8]. Y como ejemplo de áreas del conocimiento donde se ha utilizado menciona que Leyland y Goldstein (2001) proporcionan una descripción general de los modelos multinivel para la investigación en salud pública. Cressie et al. (2009) discuten modelos jerárquicos en ecología; Aitkin y Longford (1986) tienen una discusión extensa de las implicaciones prácticas de emprender un enfoque detallado del modelado jerárquico de temas controvertidos en los estudios de efectividad escolar en el Reino Unido; Sampson, Raudenbush y Earls (1997) discuten un estudio de la delincuencia utilizando un modelo jerárquico de barrios de la ciudad; y Gelman y King (1993) discuten el problema de la predicción de elecciones presidenciales con referencia a trabajos anteriores en la literatura conométrica y de las ciencias políticas. La perspectiva sobre el análisis de la varianza dada por Gelman dice que parte de su libro de 2005, y que trabajos anteriores en líneas similares incluyen a Plackett (1960), Yates (1967) y Nelder (1977, 1994), y Hodges y Sargent (2001). También menciona que Volfovsky y Hoff (2012) proponen una clase de modelos estructurados para parámetros de regresión jerárquica, que van más allá del modelo simple de coeficientes intercambiables por grupos.

Snijders y Bosker explican que hay modelos multinivel *imperfectos*, aquellos donde puede haber *cruces* en los que un individuo no necesariamente pertenece siempre al mismo grupo específico, por ejemplo, que el estudiante cambie de colegio, recibiendo efectos de ambos ambientes, o en donde se estudian dos *macro niveles* simultáneamente, por ejemplo, el efecto del colegio y del barrio. Y mencionan que el análisis de mediciones repetidas, es decir, el análisis de datos longitudinales, es otro caso especial de modelamiento multinivel. Aquel que no es *imperfecto* anotan que es el que usualmente se denomina como *modelamiento jerárquico*.

Hoff proporciona referencias a artículos seminales acerca de los modelos jerárquicos bajo el paradigma bayesaino[11]. Menciona el artículo escrito por Denis Victor Lindley y Adrian Frederick Melhuish Smith que fue publicado en 1972 por la serie metodológica del Journal of the Royal Statistical Society: *Bayes Estimates for the Linear Model*. Menciona que el efecto de shrinkage que se abordará más adelante fue expuesto en un libro publicado por Truman Lee Kelley en 1927: *The interpretation of Educational Measurement*.

Metodológicamente, el modelamiento jerárquico reconoce que es preferible modelar la realidad mediante modelos sencillos, lo que se denomina *preferencia por los modelos parsimoniosos*. Una

opción de modelamiento es considerar que agrupar los datos es irrelevante y por ende se realice un análisis tradicional. Jackman[13] lo denomina el modelamiento de *agrupación completa*. Otra opción es realizar varios análisis paralelos, uno por establecimiento educativo. Jackman lo denomina la opción de *No agrupar*. Este autor explica cómo existe una tensión entre parsimonia y realismo, entre los resultados de alto sesgo/baja varianza que podrían obtenerse con un análisis que ignora la heterogeneidad de los parámetros entre grupos, y las inferencias de bajo sesgo/alta varianza del grupo de los análisis específicos. Afirma que los modelos jerárquicos son un buen punto intermedio (pag. 302).

Gelman[6] explica cómo la estadística Bayesiana es un equivalente natural de los modelos jerárquicos, ya que considera los parámetros como variables aleatorias, no como valores fijos a ser estimados. De este modo, es posible representar un modelo jerárquico en términos Bayesianos.

Los procedimientos que se apoyan en la estadística Bayesiana interpretan los fenómenos en términos de la creencia de cuán probable es para un individuo que el evento ocurra. El grado de creencia puede basarse en el conocimiento previo sobre el evento, como conocedor de resultados de fenómenos similares anteriores, o de las creencias personales sobre el evento. Esta creencia depende de información tanto cualitativa como cuantitativa a la que tenga acceso quien ejecuta el procedimiento. Una de las características centrales de la estadística Bayesiana es que la incertidumbre sobre los valores desconocidos puede ser expresada en términos de distribuciones de probabilidad, la cual modela la creencia del individuo sobre la probabilidad de ocurrencia del fenómeno. Este conocimiento puede existir antes de observar el fenómeno, en cuyo caso la distribución se denomina *distribución previa*, o puede ser el resultado de una ponderación entre la información "previa" y la data, lo que se denomina *distribución posterior*.

El modelo jerárquico Bayesiano trata generalmente la variabilidad entre los diferentes grupos como un parámetro desconocido, generando estimaciones que se encuentran entre la gran media de todos los datos (individuos) y las medias específicas de cada grupo. Este fenómeno se denomina shrinkage en inglés, que se puede traducir como *contracción*, porque la distribución de los parámetros del nivel de grupo se *reduce* alrededor de la gran media en relación con la distribución de los parámetros del nivel de grupo que se obtendrían sin agrupación.

La teoría Bayesiana permite el modelamiento jerárquico siempre que sea justificable que las medias de los grupos sean *intercambiables*. La intercambiabilidad tiene que ver con el desconocimiento que se tiene a priori respecto a los parámetros. Si no hay conocimiento previo que indique que los parámetros de ciertos grupos deben ser diferentes a otros, podemos suponer que son *intercambiables*.

Jackman[13] explica que si los parámetros específicos del grupo son intercambiables, entonces la información sobre la media de cada grupo fluye influenciando las inferencias que se realicen acerca de los otros parámetros. La inferencia Bayesiana para cualquier media en particular reflejará una combinación de información sobre la media de dicho grupo y el componente jerárquico del modelo, generando el *shinkage*. Expresa que lo interesante de los modelos jerárquicos es que, dado que los parámetros con que se modela la media también son desconocidos, se están usando a su vez los datos para actualizar creencias previas sobre los parámetros de los niveles superiores. Como resultado, los datos de cada grupo en particular ayudan a dar forma a la densidad posterior de las distribuciones de los parámetros de los restantes grupos. Como ya se deduce de lo discutido, este tipo de *compartir* o *tomar prestada* información entre grupos es una consecuencia del modelo jerárquico, y solo tiene sentido si las medias son intercambiables.

Se presentan a continuación varios modelos Bayesianos que pueden contestar las preguntas expresadas en el planteamiento del problema, que hacen uso de las características mencionadas y que

aterrizan la teoría en una aplicación específica.

En dos de estos modelos se modela también la varianza. Hoff[11] anota en su libro que el modelamiento de la varianza es poco común, tal vez porque la media es el parámetro de mayor interés, pero opina que asumir una varianza común a todos los grupos puede conducir a una agrupación impropia de la información o a un *shrinkage* de algunos grupos en una cantidad no óptima. En este aspecto, Gelman[8] también opina que el enfoque Bayesiano difiere respecto a la literatura no Bayesiana acerca de las regresiones con coeficientes aleatorios en el hecho de promediar sobre la incertidumbre en la distribución posterior de los parámetros jerárquicos, lo cual es importante en problemas con gran incertidumbre posterior en el parámetro de la varianza (pag. 401).

# 4. Modelamiento

En el sector educativo los estudiantes se agrupan en aulas, las cuales pertenecen a una sede, la cual pertenece a un establecimiento educativo, que a su vez está inmerso en una entidad territorial. Se trata de un sector claramente jerárquico.

En el problema que se aborda se tiene información acerca una intervención que se realiza a nivel de grupos (aulas) que pertenecen a establecimientos educativos identificados, en la que se realiza la medición a estudiantes. En términos de Snijder y Bosker, el estudio *investiga proposiciones acerca de variables que están conectadas a través de diferentes niveles.*

Se consideran varios modelos. El primero, denominado *modelo base*, es el más parsimonioso. Modela la variabilidad en la media entre los establecimientos educativos, lo cual permitiría responder cuál es la probabilidad de que el valor de los índices diferencia sean mayores a cero. Pero no modela la variabilidad en la varianza, que tanto Hoff como Gelman consideran importante. En específico, Hoff afirma: "El modelamiento jerárquico de la varianza no es usual. Quizás se debe a que el parámetro de la media es el de interés. No obstante, asumir erróneamente una varianza común puede conducir a un inadecuado agrupamiento de la información, o a una contracción de los parámetros de algunos grupos específicos en una cantidad inapropiada"([11], pag 147).

El modelo base involucra dos niveles: estudiantes y establecimientos educativos.

El segundo modelo involucra la variabilidad en la media por aula y la variabilidad en la varianza por institución educativa. Es un modelo más completo. Permitirá responder cuál es la probabilidad de que el valor de los índices diferencia analizados sean diferentes de cero. Ambos modelos podrían contestar si la intervención logra efectos uniformes en todas las instituciones educativas o si se ve afectada por los contextos individuales de cada establecimiento educativo y/o aula.

El tercer modelo va dirigido a contestar si hay un efecto diferencial por sexo o por grado. Se limita a dos niveles: estudiantes y aulas o estudiantes y establecimientos educativos, ya que la brecha por sexo se puede estudiar a nivel de aula y la de grado sólo a nivel de establecimiento educativo (no hay variabilidad de grado dentro de una misma aula).

## 4.1. Modelo base: jerárquico de dos niveles

El primer modelo es un modelo Bayesiano sencillo que se usará como parámetro de comparación con los dos modelos subsecuentes y que son los de interés. Permitirá determinar si una estructura jerárquica menos parsimoniosa es valiosa en el contexto de las competencias ciudadanas y habilidades para la vida evaluadas.

Sea k el subíndice que representa cada establecimiento educativo, con $k = \{1, 2, 3, \ldots, M\}$

Sea j el subíndice que representa cada aula dentro de un establecimiento educativo, con $j = \{1, 2, 3, \ldots, n_k\}$

Sea i el subíndice que representa cada estudiante de un aula, con $i = \{1, 2, 3, \ldots, n_{jk}\}$

Su modelo es el siguiente:

$$\{Y_{ijk} \mid \theta_k, \sigma^2\} \overset{\text{ind}}{\sim} N(\theta_k, \sigma^2) \tag{1}$$

Obsérvese que $\theta_k$ sólo cuenta con un subíndice que representa a la institución educativa, por tanto,

es el mismo para todas las aulas. $k = \{1, 2, 3, \ldots, M\}$. El parámetro $\sigma^2$ se supone igual para todas las instituciones educativas. La parsimonia no espera que la variabilidad en el índice entre las aulas de una misma institución educativa sea tal que valga la pena modelarlos individualmente.

No se tiene información a priori que indique cómo unos establecimientos educativos son diferentes a los restantes. Podemos suponer de una manera razonable la *intercambiabilidad* de la función de probabilidad marginal de los índices diferencia que se modelan para cada grupo de estudiantes. Como aduce Hoff[11], si los grupos mismos son muestras de alguna población de grupos, entonces el supuesto de intercambiabilidad de los parámetros específicos de cada grupo pueden ser apropiados (pag. 131).

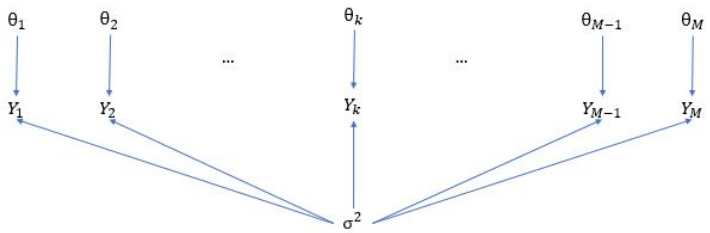



Figura 1: Modelo base

En la Figura 1 $Y_k$ representa el vector de resultados de un establecimiento educativo cualquiera k, que cuenta con $\sum_{j=1}^{n_k} n_{jk}$ estudiantes.

Como *resultado* se utiliza el índice diferencia. No se incorpora en el modelo el resultado de la prueba previa a la intervención y aquella realizada una vez terminada la intervención, sino directamente la diferencia como ya se explicó en el capítulo 2.

### 4.1.1 Distribución de las previas

Sea la distribución previa del primer parámetro de la ecuación (1)

$$\{\theta_k \mid \mu_0, \tau_0^2\} \overset{\text{ind}}{\sim} N(\mu_0, \tau_0^2) \tag{2}$$

Con esto se está modelando el segundo nivel de la estructura jerárquica. Es decir, cada valor esperado de la institución educativa específica k se modela como una variable aleatoria de distribución normal.

Sea la distribución previa del segundo parámetro de la ecuación (1):

$$\{\sigma^2 \mid y_0, z_0\} \overset{\text{ind}}{\sim} IGamma(y_0, z_0), \tag{3}$$

IGamma porque modela números positivos y la varianza está en dicho dominio de los números reales.

$\mu_0$, $\tau_0^2$, $y_0$ y $z_0$ son hiperparámetros que se definen a priori.

### 4.1.2 Inferencia de las distribuciones posteriores

La inferencia de las distribuciones conjuntas posteriores conjugadas de los parámetros llevan a los siguientes resultados (en el Anexo C se desarrolla la formulación para llegar a estos resultados):

$$p(\theta_k \mid \sigma^2, y_{ijk}, \mu_0, \tau_0^2) \sim N\Big(\Big[\frac{1}{\tau_0^2}\mu_0 + \frac{\sum_{j=1}^{n_k} n_{jk}}{\sigma^2}\bar{y}_k\Big]\Big[\frac{\tau_0^2\sigma^2}{\sigma^2 + \sum_{j=1}^{n_k} n_{jk}\tau_0^2}\Big], \Big[\frac{\tau_0^2\sigma^2}{\sigma^2 + \sum_{j=1}^{n_k} n_{jk}\tau_0^2}\Big]\Big) \quad (4)$$

$$p(\sigma^2 \mid \theta_k, y_{ijk}, y_0, z_0) \sim IGamma\Big(y_0 + \frac{1}{2}\sum_{k=1}^{M}\sum_{j=1}^{n_k} n_{jk}, z_0 + \frac{1}{2}\sum_{k=1}^{M}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\big(y_{ijk} - \theta_k\big)^2\Big) \quad (5)$$

Se debe entender que $\sum_{j=1}^{n_k} n_{jk}$ son todos los estudiantes de la institución educativa k,

que $\sum_{k=1}^{M}\sum_{j=1}^{n_k} n_{jk}$ es el total de estudiantes de todas las instituciones educativas,

y que $\sum_{k=1}^{M}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\left(y_{ijk} - \theta_k\right)^2$ es la suma sobre la totalidad de instituciones educativas, del cuadrado de la desviación de los puntajes de los alumnos, respecto a la media de su respectiva institución educativa.

## 4.2 Segundo modelo: jerárquico de tres niveles

Se parte de un modelo de este tipo:

$$\{Y_{ijk} \mid \theta_{jk}, \sigma_k^2\} \overset{\text{ind}}{\sim} N(\theta_{jk}, \sigma_k^2) \quad (6)$$

Obsérvese que el subíndice de $\theta_{jk}$ implica que es diferente para cada aula, por tanto, habrá tantos parámetros $\theta$ como aulas. El parámetro $\sigma_k^2$ supone igual varianza para todas las aulas de una misma institución educativa, por eso sólo cuenta con el subíndice k. No se espera que la variabilidad en el índice entre las aulas de una misma institución educativa sea tal que valga la pena modelarlos individualmente.

La discusión acerca del supuesto de intercambiabilidad dada para el modelo base también es válida para este modelo y los siguientes.

La Figura 3 es muy relevante para entender el modelo. Léase el texto observando simultáneamente ésta para una más rápida comprensión. La Figura 2 es un paso intermedio para ayudar a la explicación.

### 4.2.1 Distribuciones de las previas

Sea la distribución previa del primer parámetro de la ecuación (6)

$$\{\theta_{jk} \mid \mu_k, \tau_k^2\} \overset{\text{ind}}{\sim} N(\mu_k, \tau_k^2) \quad (7)$$

Con esto se está modelando el segundo nivel de la estructura jerárquica. Es decir, cada valor esperado del aula específica j de la institución educativa específica k se modela como una variable aleatoria.

La ecuación (6) representa la heterogeneidad intra-grupos. La ecuación (7) representa la heterogeneidad entre-grupos.

Ahora, sea la distribución previa del segundo parámetro de la ecuación (6):

$$\{\sigma_k^2 \mid \alpha, \eta\} \overset{\text{ind}}{\sim} IGamma(\alpha, \eta), \tag{8}$$

IGamma porque modela números positivos y la varianza está en dicho dominio de los números reales, pero además, porque la posterior es conjugada de la previa.

Con esta especificación, se termina de modelar el nivel aulas. (Ver Figura 2)

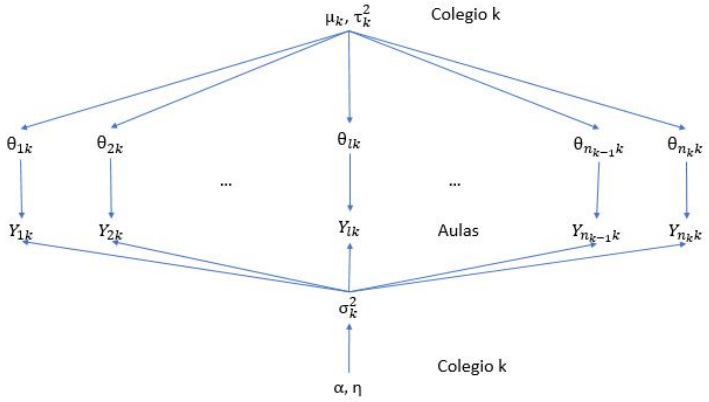



Figura 2: Niveles 1 y 2 del segundo modelo, jerárquico de tres niveles, para una institución educativa cualquiera k.

Sea la distribución previa del primer parámetro de la ecuación (7)

$$\{\mu_k \mid \gamma, \kappa^2\} \overset{\text{iid}}{\sim} N(\gamma, \kappa^2) \tag{9}$$

Se está modelando el tercer nivel de la estructura jerárquica. Así que el valor esperado de cada institución educativa k se modela como una variable aleatoria.

La distribución previa del segundo parámetro de la ecuación (7), $\tau_k^2$, se modela también como una IGamma:

$$\{\tau_k^2 \mid \lambda, \xi\} \sim IGamma(\lambda, \xi) \tag{10}$$

Sea modelado ahora el primer parámetro de la ecuación (8). Hoff [11] presenta que el parámetro de forma de una distribución gamma inversa, *restringiéndolo a ser un número entero*, se puede modelar con una distribución previa exponencial tal que:

$$p(\alpha) \propto e^{-\alpha a_0}, \tag{11}$$

Y para el segundo parámetro, $\eta$, se puede definir una previa tal que

$$p(\eta) \propto Gamma(b_0, c_0), \tag{12}$$

$a_0$, $b_0$ y $c_0$ son hiperparámetros que se definen a priori.

El modelado del tercer nivel finaliza especificando las distribuciones previas de los parámetros $\gamma$ y $\kappa^2$ de la ecuación (9):

$$\gamma \sim N(g_0, h_0^2) \tag{13}$$

$$\kappa^2 \sim IGamma(u_0, v_0), \tag{14}$$

donde $g_0$, $h_0$, $u_0$ y $v_0$ son hiperparámetros.

Los parámetros $\lambda$ y $\xi$ de la ecuación (10) también se pueden especificar de la misma forma que se presentan en las ecuaciones (11) y (12):

$$p(\lambda) \propto e^{-d_0 \lambda}, \tag{15}$$

$$p(\xi) \propto Gamma(e_0, f_0) \tag{16}$$

$d_0$, $e_0$ y $f_0$ son hiperparámetros.

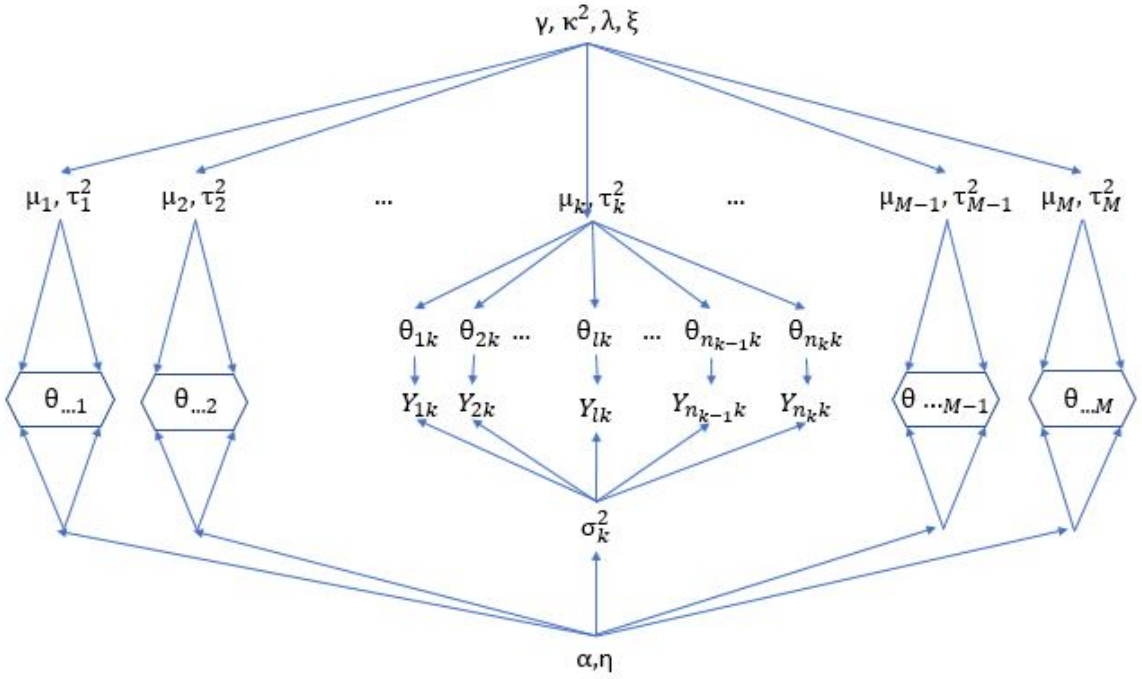



Figura 3: Nivel 1, 2 y 3 del modelo jerárquico de tres niveles (con detalle de los niveles 1 y 2 de una institución educativa cualquiera k)

Esta especificación implica estimar aproximadamente $\sum_{k=1}^{M} n_k$ parámetros $\theta$, y M parámetros de cada $\sigma^2$, $\mu$ y $\tau^2$; y un único parámetro de: $\alpha$, $\eta$, $\gamma$, $\kappa^2$, $\lambda$ y $\xi$. Un total de 3M + $\sum_{k=1}^{M} n_k$ + 6 parámetros.

También se requiere definir para el nivel 2 seis hiperparámetros $\{a_0, b_0, c_0, d_0, e_0, f_0\}$, y cuatro para el nivel 3: $\{g_0, h_0, u_0, v_0\}$.

La Figura 3 sólo representa las variables aleatorias. No los hiperpartámetros.

Resumen:

| **Parámetro** | **Previa** |
|---|---|
| $\theta_{jk}$ | $\{\theta_{jk} \mid \mu_k, \tau_k^2\} \overset{\text{ind}}{\sim} N(\mu_k, \tau_k^2)$ |
| $\sigma_k^2$ | $\{\sigma_k^2 \mid \alpha, \eta\} \overset{\text{ind}}{\sim} IGamma(\alpha, \eta)$ |
| $\mu_k$ | $\{\mu_k \mid \gamma, \kappa^2\} \overset{\text{iid}}{\sim} N(\gamma, \kappa^2)$ |
| $\tau_k^2$ | $\{\tau_k^2 \mid \lambda, \xi\} \overset{\text{iid}}{\sim} IGamma(\lambda, \xi)$ |
| $\alpha$ | $p(\alpha \mid a_0) \propto e^{-\alpha a_0}$ |
| $\eta$ | $p(\eta \mid b_0, c_0) \propto Gamma(b_0, c_0)$ |
| $\gamma$ | $\gamma \sim N(g_0, h_0^2)$ |
| $\kappa^2$ | $\kappa^2 \sim IGamma(u_0, v_0)$ |
| $\lambda$ | $\lambda \propto e^{-\lambda d_0}$ |
| $\xi$ | $\xi \sim Gamma(e_0, f_0)$ |

Cuadro 3: Resumen Previas del modelo jerárquico de tres niveles.

### 4.2.2 Inferencia de las distribuciones posteriores

La inferencia de las distribuciones conjuntas posteriores conjugadas de los parámetros llevan a los siguientes resultados (En el Anexo A se desarrolla la formulación para llegar a estos resultados):

$$p(\theta_{jk} \mid \sigma_k^2, y_{ijk}, \mu_k, \tau_k^2) \sim N\Big(\Big[\frac{1}{\tau_k^2}\mu_k + \frac{n_{jk}}{\sigma_k^2}\bar{y}_{jk}\Big]\Big[\frac{\tau_k^2\sigma_k^2}{\sigma_k^2 + n_{jk}\tau_k^2}\Big], \Big[\frac{\tau_k^2\sigma_k^2}{\sigma_k^2 + n_{jk}\tau_k^2}\Big]\Big) \tag{17a}$$

$$p(\sigma_k^2 \mid \theta_{jk}, y_{ijk}, \alpha, \eta) \sim IGamma\Big(\alpha + \frac{1}{2}\sum_{j=1}^{n_k} n_{jk}, \eta + \frac{1}{2}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\Big(y_{ijk} - \theta_{jk}\Big)^2\Big) \tag{17b}$$

$$p(\mu_k \mid \gamma, \theta_{jk}, \kappa^2, \tau_k^2) \sim N\left(\left[\frac{1}{\kappa^2}\gamma + \frac{n_k}{\tau_k^2}\bar{\theta}_{jk}\right]\left[\frac{\kappa^2\tau_k^2}{\tau_k^2 + n_k\kappa^2}\right], \left[\frac{\kappa^2\tau_k^2}{\tau_k^2 + n_k\kappa^2}\right]\right) \tag{17c}$$

$$p(\tau_k^2 \mid \lambda, \xi, \theta_{jk}, \mu_k) \sim IGamma\Big(\lambda + \frac{n_k}{2}, \xi + \frac{1}{2}\sum_{j=1}^{n_k}(\theta_{jk} - \mu_k)^2\Big) \tag{17d}$$

$$p(\alpha \mid \sigma_k^2, \eta) \propto \frac{\eta^{(M\alpha)}e^{(-\alpha a_0)}}{\Gamma(\alpha)^M}\prod_{k=1}^{M}(\sigma_k^2)^{-(\alpha+1)} \tag{17e}$$

$$p(\eta \mid \alpha, \sigma_k^2) \sim Gamma\left(b_0 + M\alpha, c_0 + \sum_{k=1}^{M}\frac{1}{\sigma_k^2}\right) \tag{17f}$$

$$p(\gamma \mid \kappa, \mu_k) \sim N\left(\left[\frac{1}{h_0^2}g_0 + \frac{M}{\kappa^2}\bar{\mu}_k\right]\left[\frac{h_0^2\kappa^2}{\kappa^2 + Mh_0^2}\right], \left[\frac{h_0^2\kappa^2}{\kappa^2 + Mh_0^2}\right]\right) \tag{17g}$$

$$p(\kappa^2 \mid \mu_k, \gamma) \sim IGamma\Big(u_0 + \frac{M}{2}, v_0 + \frac{1}{2}\sum_{k=1}^{M}(\mu_k - \gamma)^2\Big) \tag{17h}$$

$$p(\lambda \mid \tau_k^2, \xi) \sim \frac{\xi^{(M\lambda)}e^{(-\lambda d_0)}}{\Gamma(\lambda)^M}\prod_{k=1}^{M}(\tau_k^2)^{-(\lambda+1)} \tag{17i}$$

$$p(\xi \mid \lambda, \tau_k^2) \sim Gamma\left(e_0 + M\lambda, f_0 + \sum_{k=1}^{M}\frac{1}{\tau_k^2}\right) \tag{17j}$$

### 4.3 Tercer modelo: jerárquico de dos niveles considerando cofactores

Se puede plantear un análisis con cofactores. Ya se mencionó que en el sector educativo interesa, por ejemplo, detectar si hay brecha entre sexos, o si la implementación de una intervención es más conveniente en unos grados (edades) que en otros. Sexo o grado son cofactores. Son variables categóricas que se incorporan al modelo como variables dummy. La sección 4.3.3 explica de qué manera se manejan e interpretan.

El planteamiento tiene mucho en común con el modelo jerárquico de tres niveles, la diferencia radica en que este modelo incorpora covariables en la distribución de muestreo (verosimilitud) y se parece al planteamiento de un modelo de regresión lineal mixto, jerárquico:

$$y_{ijk} = \beta_{jk}^{(0)} + \beta_{jk}^{(1)} x_{ijk}^{(1)} + \cdots + \beta_{jk}^{(p)} x_{ijk}^{(p)} + \varepsilon_{ijk} \tag{18}$$

con

$$\varepsilon_{ijk} \sim N(0, \sigma_k^2) \tag{19}$$

Se propone el modelo:

$$\{y_{ijk} \mid \boldsymbol{\beta}_{jk}, \boldsymbol{x}_{ijk}, \sigma_k^2\} \overset{\text{ind}}{\sim} N_{p+1}(\boldsymbol{X}_{ijk}^T \boldsymbol{\beta}_{jk}, \sigma_k^2), \tag{20}$$

muy similar al modelo jerárquico de tres niveles.

Sea en el planteamiento actual la multiplicacióin de $\boldsymbol{X}_{ijk}^T \boldsymbol{\beta}_{jk}$ el equivalente al parámetro $\theta_{jk}$, donde $\boldsymbol{\beta}_{jk} = (\beta_{jk}^{(0)}, \beta_{jk}^{(1)} \ldots, \beta_{jk}^{(p)})$, $\boldsymbol{X}_{ijk} = (x_{ijk}^{(0)}, x_{ijk}^{(1)}, \ldots, x_{ijk}^{(p)})$ y $p$ el número de covariables.

Obsérvese que se está utilizando el subíndice $jk$ para representar los cofactores a nivel de aula. Si se representara a nivel de establecimiento educativo, la formulación es:

$$\{y_{ijk} \mid \boldsymbol{\beta}_{k}, \boldsymbol{x}_{ijk}, \sigma_k^2\} \overset{\text{ind}}{\sim} N_{p+1}(\boldsymbol{X}_{ijk}^T \boldsymbol{\beta}_{k}, \sigma_k^2), \tag{21}$$

#### 4.3.1 Distribuciones previas

Se modela la distribución previa del vector $\boldsymbol{\beta}_{jk}$ como una normal multivariada de orden $p+1$:

$$\{\boldsymbol{\beta}_{jk} \mid \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k\} \sim N_{p+1}(\boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) \tag{22}$$

Si se representara a nivel de establecimiento educativo, la formulación es:

$$\{\boldsymbol{\beta}_{k} \mid \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k\} \sim N_{p+1}(\boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) \tag{23}$$

De manera general,

$$\boldsymbol{\psi}_k \sim N_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \tag{24}$$

$$\boldsymbol{\Sigma}_k \sim IWishart(\nu_0, \mathbf{S}_0);\ \nu_0, \mathbf{S}_0 > 0 \tag{25}$$

y la varianza del término aleatorio como

$$\sigma_k^2 \sim IG(\alpha, \gamma) \tag{26}$$

donde se observa que tanto el vector de medias como la matriz de varianza se modela como variables aleatorias, del mismo modo a como se realizó en el modelo jerárquico de tres niveles por medio de las ecuaciones (7) y (8). La distribución Wishart inversa es el equivalente multivariado de la distribución Gamma inversa.

Los restantes componentes del modelos son:

$$p(\alpha) \propto e^{-\alpha a_0}, \tag{27}$$

$$p(\gamma) \propto Gamma(b_0, c_0), \tag{28}$$

El valor $a_0$, $b_0$, $c_0$ y $\nu_0$ y las matrices $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$ y $\mathbf{S}_0$ son hiperparámetros.

### 4.3.2 Inferencia de las distribuciones posteriores.

La inferencia de las distribuciones conjuntas posteriores conjugadas de los parámetros llevan a los siguientes resultados (en el Anexo B se desarrolla la formulación para llegar a estos resultados):

**Nivel aula**:

$$p(\boldsymbol{\beta}_{jk} \mid \boldsymbol{\Sigma}_k, \sigma_k, \boldsymbol{\psi}_k, y_{ijk}) \sim N\Bigg( \Big(\boldsymbol{\Sigma}_k^{-1} + \sigma_k^{-2} \sum_{i=1}^{n_{jk}} \boldsymbol{X}_{ijk}\boldsymbol{X}_{ijk}^T\Big)^{-1} \Big(\boldsymbol{\Sigma}_{\boldsymbol{k}}^{-\mathbf{1}}\boldsymbol{\psi}_k + \sigma_k^{-2} \sum_{i=1}^{n_{jk}} (y_{ijk}\boldsymbol{X}_{ijk}^T)\Big), \Big(\boldsymbol{\Sigma}_k^{-1} + \sigma_k^{-2} \sum_{i=1}^{n_{jk}} \boldsymbol{X}_{ijk}\boldsymbol{X}_{ijk}^T\Big)^{-1} \Bigg) \tag{29}$$

$$(\boldsymbol{\psi}_k \mid \boldsymbol{\Sigma}_{\boldsymbol{k}}, \boldsymbol{\beta}_{jk}) \sim N\Big( (\boldsymbol{\Lambda}^{-1} + n_k\boldsymbol{\Sigma}_k^{-1})^{-1}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{.k} + \frac{1}{2}\boldsymbol{\Lambda}^{-1}\boldsymbol{\mu}), (\boldsymbol{\Lambda}^{-1} + n_k\boldsymbol{\Sigma}_k^{-1})^{-1} \Big) \tag{30}$$

$$(\boldsymbol{\Sigma}_k \mid \boldsymbol{\psi}_k, \boldsymbol{\beta}_{jk}) \sim IWishart\Big( \nu_0 + n_k, \Big(S_0 + \sum_{j=1}^{n_k} (\boldsymbol{\beta}_{jk} - \boldsymbol{\psi}_k)(\boldsymbol{\beta}_{jk} - \boldsymbol{\psi}_k)^T\Big) \Big) \tag{31}$$

$$(\sigma_k^2 \mid y_{ijk}, \boldsymbol{\beta}_{jk}) \sim IG\Big( \alpha + \frac{1}{2}N_k, \gamma + \frac{1}{2}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}} (y_{ijk} - \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk})^2 \Big) \tag{32}$$

$$(\gamma \mid \sigma_k^2) \sim Gamma\Big( b_0 + M\alpha, c_0 + \sum_{k=1}^{M} \frac{1}{\sigma_k^2} \Big) \tag{33}$$

$$\alpha \sim e^{-\alpha a_0} \tag{34}$$

Con $N_k = \sum_{j=1}^{n_k} n_{jk}$

Esta especificación implica estimar aproximadamente $\sum_{j=1}^{M} n_k$ parámetros de $\boldsymbol{\beta}$, que es un vector de tamaño $p+1$ y M parámetros de: $\boldsymbol{\Psi}_k$, $\boldsymbol{\Sigma}_k$ y $\sigma_k^2$, Las dos primeras son matrices cuyos tamaños se presentan más adelante. Un total de $\sum_{j=1}^{M} n_k + M \times (p+1) + M \times (p+1) \times (p+1)$ + M + 2 parámetros.

**Nivel establecimiento educativo**:

$$p(\boldsymbol{\beta}_k \mid \boldsymbol{\Sigma}_k, \sigma_k, \boldsymbol{\psi}_k, y_{ijk}) \sim N\Bigg( \Big( \boldsymbol{\Sigma}_k^{-1} + \sigma_k^{-2} \sum_{j=1}^{n_k} \sum_{i=1}^{n_{jk}} (\boldsymbol{X}_{ijk} \boldsymbol{X}_{ijk}^T) \Big)^{-1} \Big( \boldsymbol{\Sigma}_{\boldsymbol{k}}^{-\mathbf{1}} \boldsymbol{\psi}_k + \sigma_k^{-2} \sum_{j=1}^{n_k} \sum_{i=1}^{n_{jk}} (y_{ijk} \boldsymbol{X}_{ijk}^T) \Big), \Big( \boldsymbol{\Sigma}_k^{-1} + \sigma_k^{-2} \sum_{j=1}^{n_k} \sum_{i=1}^{n_{jk}} (\boldsymbol{X}_{ijk} \boldsymbol{X}_{ijk}^T) \Big)^{-1} \Bigg) \tag{35}$$

$$(\boldsymbol{\psi}_k \mid \boldsymbol{\Sigma}_k, \boldsymbol{\beta}_k) \sim N\Big( (\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Sigma}_k^{-1})^{-1} (\boldsymbol{\Sigma}_k^{-1} \boldsymbol{\beta}_k + \boldsymbol{\Lambda}^{-1} \boldsymbol{\mu}), (\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Sigma}_k^{-1})^{-1} \Big) \tag{36}$$

$$(\boldsymbol{\Sigma}_k \mid \boldsymbol{\Sigma}_k, \boldsymbol{\psi}_k, \boldsymbol{\beta}_k) \sim IWishart\Big( \nu_0 + (p-1), S_0 + (\boldsymbol{\beta}_k - \boldsymbol{\psi}_k)(\boldsymbol{\beta}_k - \boldsymbol{\psi}_k)^T \Big) \tag{37}$$

$$(\sigma_k^2 \mid y_{ijk}, \boldsymbol{\beta}_k) \sim IG\Big( \alpha + \frac{1}{2} N_k, \gamma + \frac{1}{2} \sum_{j=1}^{n_k} \sum_{i=1}^{n_{jk}} (y_{ijk} - \boldsymbol{X}_{ijk}^T \boldsymbol{\beta}_k)^2 \Big) \tag{38}$$

$$(\gamma \mid \sigma_k^2) \sim Gamma\Big( b_0 + M\alpha, c_0 + \sum_{k=1}^{M} \frac{1}{\sigma_k^2} \Big) \tag{39}$$

$$\alpha \sim e^{-\alpha a_0} \tag{40}$$

Con $N_k = \sum_{j=1}^{n_k} n_{jk}$

Esta especificación implica estimar aproximadamente M parámetros de: $\boldsymbol{\beta}$, $\boldsymbol{\Psi}_k$, $\boldsymbol{\Sigma}_k$ y $\sigma_k^2$. Las tres primeras son matrices cuyo tamaño se especifica en la siguiente sección. Un total de $2 * M \times (p+1)$ + $M \times (p+1) \times (p+1)$ + M + 2 parámetros.

### 4.3.3 Dimensión de las matrices

En el planteamiento del problema se expuso por qué interesan dos cofactores: el sexo y el grado. El sexo es una una variable indicadora o dummy binaria, por tanto es una covariable. Los grados educativos del problema en que se aplica el modelo se transforman en dummies, resultando en el número de grados, menos uno, covariables.

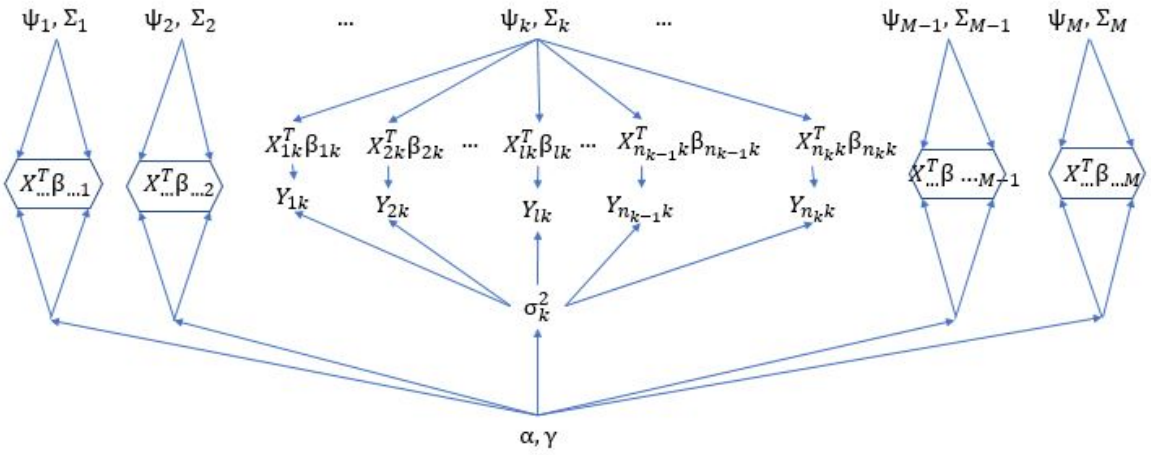



Figura 4: Modelos jerárquico de dos niveles, a nivel de aula, con detalle de una institución educativa cualquiera k.

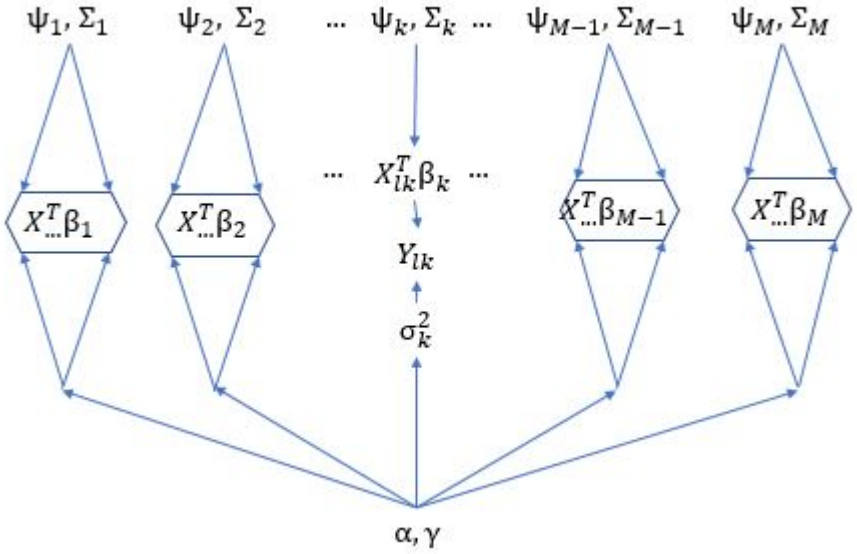



Figura 5: modelos jerárquicos de dos niveles, a nivel de establecimiento educativo, con el detalle de una institución educativa cualquiera k.

Una opción es incluir simultáneamente todas las covariables, en consecuencia, $p+1 = 5$. Recuérdese que el término del intercepto suma uno adicional.

La interpretación es la misma que en los modelos de regresión lineal.

Por ejemplo, $\beta_{jk}^{(0)}$ es el valor medio de la contribución del sexo masculino en grado sexto a la variable $y$ del aula j en el establecimiento educativo k.

$\beta_{jk}^{(0)} + \beta_{jk}^{(1)}$ es el valor medio de la contribución del sexo femenino en el grado de referencia a la variable $y$ del aula j en el establecimiento educativo k. Pero en el aula j, todos tienen el mismo grado. No hay variabilidad. Se podrían platear dos opciones:

- Evaluar sexo y grado a nivel de establecimiento educativo.
- Realizar dos análisis, uno a nivel de aula para sexo y otro a nivel de establecimiento educativo para grado.

Si se evaluaran sexo y grado como covariables simultáneas, el coeficiente $\beta_k^{(0)}$ representa combinadamente el sexo masculino en el grado de referencia, impidiendo realizar un análisis del aporte del sexo independientemente del grado, o del aporte del grado independientemente del sexo. Desde el punto de vista de política pública interesan estas preguntas por separado, por tanto se realizarán los dos análisis por separado. Se denomina primer modelo jerárquico de dos niveles el que atañe al sexo, en el que p = 1; y segundo modelo jerárquico de dos niveles al que atañe al grado, con un p = 3, ya que los datos que se tienen corresponden a intervenciones realizadas en grados sexto a noveno..

Entonces, de manera específica, a **nivel de aula**

$$\underset{\sum n_{ijk}\times 1}{y_{ijk}} \sim N(\underset{p+1\times\sum n_{ijk}}{\boldsymbol{X_{ijk}^T}}\underset{\sum n_{jk}\times(p+1)}{\boldsymbol{\beta_{jk}}}, \underset{M\times 1}{\sigma_k^2})$$

M es el número total de instituciones educativas, $\sum n_{jk}$ el número de aulas y $\sum n_{ijk}$ el número de estudiantes.

$\boldsymbol{X}$ es una matriz de $\sum n_{ijk} \times (p+1)$, donde cada fila representa la observación multivariada de un individuo. En el primer modelo jerárquico de dos niveles, la columna $\boldsymbol{x}_{ijk}^{(1)}$ contiene los datos del sexo, y $\boldsymbol{x}_{ijk}^{(0)}$ es un vector de unos que se corresponde con la constante.

En consecuencia, para un k fijo:

**Previas**

$$(\underset{n_{jk}\times(p+1)}{\boldsymbol{\beta}_{jk}} \mid \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) \sim N(\underset{(p+1)\times 1}{\boldsymbol{\psi}_k}, \underset{(p+1)\times(p+1)}{\boldsymbol{\Sigma}_k})$$

$$\underset{(p+1)}{\boldsymbol{\psi}_k} \sim N(\underset{(p+1)\times 1}{\boldsymbol{\mu}}, \underset{(p+1)\times(p+1)}{\boldsymbol{\Lambda}})$$

$$\underset{(p+1)\times(p+1)}{\boldsymbol{\Sigma}_k} \sim IWishart(\underset{1\times 1}{\nu_0}, \underset{(p+1)\times(p+1)}{\mathbf{S}_0})$$

**Posteriores**

$$\underset{n_{jk}\times(p+1)}{\boldsymbol{\beta}_{jk}} \sim N\Bigg(\Big(\underset{(p+1)\times(p+1)}{\boldsymbol{\Sigma}_k^{-1}} + \sigma_k^{-2}\sum_{i=1}^{n_{jk}} \underset{(p+1)\times(p+1)}{\boldsymbol{X}_{ijk}\boldsymbol{X}_{ijk}^T}\Big)^{-1}\Big(\underset{1\times(p+1)}{\boldsymbol{\psi_k}\boldsymbol{\Sigma_k}^{-1}} + \sigma_k^{-2}\sum_{i=1}^{n_{jk}} \underset{1\times(p+1)}{(y_{ijk}\boldsymbol{X}_{ijk}{}^T)}\Big),$$
$$\Big(\underset{(p+1)\times(p+1)}{\boldsymbol{\Sigma}_k^{-1}} + \sigma_k^{-2}\sum_{i=1}^{n_{jk}} \underset{(p+1)\times(p+1)}{\boldsymbol{X}_{ijk}\boldsymbol{X}_{ijk}^T}\Big)^{-1}\Bigg)$$

$$\underset{(p+1)}{\boldsymbol{\psi}_k} \sim N(\underset{(p+1)}{\boldsymbol{m_k}}, \underset{(p+1)\times(p+1)}{\boldsymbol{V}_k})$$
$$\underset{(p+1)\times(p+1)}{\boldsymbol{\Sigma}_k} \sim IWishart\Big(\underset{1}{\nu_k}, \underset{(p+1)\times(p+1)}{\boldsymbol{S}_k}\Big)$$

En el segundo modelo jerárquico de dos niveles, las columnas $\boldsymbol{x}_{ijk}^{(1:3)}$ contienen los datos del grado y $\boldsymbol{x}_{ijk}^{(0)}$ es un vector de unos.

Para un k fijo:

$$\underset{N_k\times 1}{y_{ijk}} \sim N(\underset{N_k\times(p+1)}{\boldsymbol{X_{ijk}^T}}\underset{(p+1)\times 1}{\boldsymbol{\beta_k}}, \underset{1}{\sigma_k^2})$$

**Previas**

$$(\underset{(p+1)\times 1}{\boldsymbol{\beta}_k} \mid \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) \sim N(\underset{(p+1)\times 1}{\boldsymbol{\psi}_k}, \underset{(p+1)\times(p+1)}{\boldsymbol{\Sigma}_k})$$

$$\underset{(p+1)}{\boldsymbol{\psi}_k} \sim N(\underset{(p+1)\times 1}{\boldsymbol{\mu}}, \underset{(p+1)\times(p+1)}{\boldsymbol{\Lambda}})$$

$$\underset{(p+1)\times(p+1)}{\boldsymbol{\Sigma}_k} \sim IWishart(\underset{1\times 1}{\nu_0}, \underset{(p+1)\times(p+1)}{\mathbf{S}_0})$$

**Posteriores**

$$\underset{(p+1)\times 1}{\boldsymbol{\beta}_k} \sim N\Bigg(\Big(\underset{(p+1)\times(p+1)}{\boldsymbol{\Sigma}_k^{-1}} + \sigma_k^{-2}\underset{(p+1)\times(p+1)}{\boldsymbol{X}_{ijk}\boldsymbol{X}_{ijk}^T}\Big)^{-1}\Big(\underset{1\times(p+1)}{\boldsymbol{\psi_k}\boldsymbol{\Sigma_k}^{-1}} + \sigma_k^{-2}\underset{1\times(p+1)}{(y_{ijk}\boldsymbol{X}_{ijk}{}^T)}\Big),$$
$$\Big(\underset{(p+1)\times(p+1)}{\boldsymbol{\Sigma}_k^{-1}} + \sigma_k^{-2}\underset{(p+1)\times(p+1)}{\boldsymbol{X}_{ijk}\boldsymbol{X}_{ijk}^T}\Big)^{-1}\Bigg)$$

$$\underset{(p+1)\times 1}{\boldsymbol{\psi}_k} \sim N(\underset{(p+1)\times 1}{m_k}, \underset{(p+1)\times(p+1)}{\boldsymbol{V}_k})$$
$$\underset{(p+1)\times(p+1)}{\boldsymbol{\Sigma}_k} \sim IWishart\Big(\underset{1}{\nu_k}, \underset{(p+1)\times(p+1)}{\boldsymbol{S}_k}\Big)$$

## 4.4 Validación de la implementación de los modelos

La distribución posterior de los modelos definidos puede ser estimada muestreando por medio de métodos de Monte Carlo (MC) mediante cadenas de Markov, (MCMC) por sus siglas en inglés, por cuanto todas las distribuciones posteriores de los modelos planteados son conjugadas de las distribuciones previas. Estas muestras se pueden usar para evaluar una integral sobre esa variable, como su valor esperado o varianza.

En la práctica se desarrollan un conjunto de vectores o *cadenas* de valores, a partir de un conjunto de puntos elegidos arbitrariamente y suficientemente separados entre sí. Estas cadenas son procesos estocásticos de *paseos aleatorios* que se mueven en un espacio de dimensión p de acuerdo con un algoritmo que busca puntos con una contribución razonablemente alta a la integral, en una sucesión que asigna cada vez mayores probabilidades a las ubicaciones.

Estos muestreadores fueron implementados en R, programados. Si bien existen paquetes que permiten calcular muchos modelos Bayesianos, éstos no admiten mucha personalización o control de lo que ocurre por dentro por cuanto están programados en C++ con el objeto de optimizar el tiempo de cómputo.

### 4.4.1 Convergencia, tamaño efectivo y error estándar

Para chequear el MCMC de cada parámetro estimado, se verifica que se haya logrado convergencia en los resultados.

La convergencia se suele verificar mediante el cálculo de la Log-verosimilitud del MCMC:

$$logP\left(\tilde{y} \mid \theta^{(s)}, (\sigma^2)^{(s)}\right) = \Sigma_{k=1}^{M}\Sigma_{j=1}^{n_k}\Sigma_{i=1}^{n_{jk}} N(\theta_{jk}, \sigma_k^2)$$

Dónde el superíndice (s) representa la iteración que corresponde.

$$(s) \in \{1, 2, 3, \ldots, S\}$$

S es el número total de iteraciones.

Gráficamente, el vector logP debe oscilar alrededor de un valor si ha alcanzado convergencia.

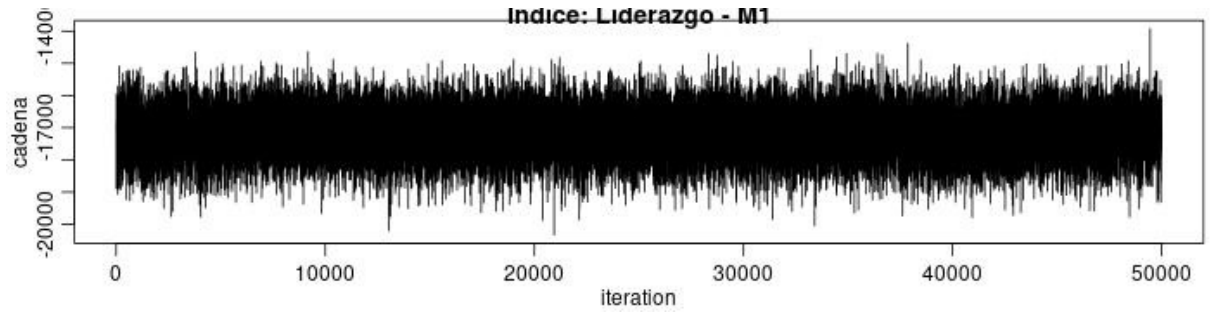



Figura 6: Gráfica de la convergencia de la log-verosimilitud

Una vez comprobada la convergencia, se deben calcular los errores estándar (EE). Una aproximación a éstos es dividir la desviación estándar entre la raíz cuadrada del tamaño efectivo de la cadena generada por el MCMC:

$$EE_{MCMC}(\gamma) = \sqrt{\frac{Var_{MCMC}(\gamma)}{S_{ef}}} \tag{41}$$

Mientras que las muestras aleatorias del integrando utilizado en una integración convencional por métodos de Monte Carlo son estadísticamente independientes, las utilizadas en MCMC están auto-correlacionadas ya que cada iteración depende del valor del paso anterior. Por esa razón se calcula un *tamaño efectivo*.

Sea $\gamma$ el parámetro estimado y sea $S_{ef}$ el *tamaño de muestra efectivo*, que también es una estimación. Se puede interpretar como el número de muestras de MCMC necesarias para dar la misma precisión que el modelo de Monte Carlo sin autocorrelaciones.

Computacionalmente el *tamaño de muestra efectivo* se obtiene mediante el cálculo:

$$S_{ef} = \frac{S}{\sum_{k=1}^{\infty} \rho_k} \tag{42}$$

donde S es el tamaño de la muestra y $\rho_k$ es la autocorrelación en el rezago k. La suma va hasta $\infty$, pero el estándar computacional es seleccionar un punto terminal, t, tal que la suma de la autocorrelación en dos rezagos sucesivos $\rho_{t-1} + \rho_t$ sea inferior a un umbral ($\epsilon$). Es de advertir que el tamaño de muestra S no coincide con el número de iteraciones sobre el cual se corrió el modelo ya que se elimina un conjunto inicial de resultados que corresponden a los obtenidos en el tiempo en que el MCMC tiende a la estabilización. Se denominan *iteraciones de calentamiento* (burn en inglés). Además para disminuir la auto-correlación propia de una cadena de Markov sólo se guardan los resultados cada cierto número de iteraciones, 5, 10 ó 20, dependiendo de cada modelo.

El error estándar (EE) debe ser pequeño respecto al valor estimado del parámetro. Se usa que $\bar{\gamma}/EE(\gamma) > 2$, representando $\gamma$ cualquier parámetro.

A mayor autocorrelación, menor tamaño efectivo de muestra y, por ende, mayor error estándar.

En el numeral 6.6 del libro de Hoff[11] se halla una introducción a la validación de MCMC.

### 4.4.2 Posterior predictive p-value (ppp)

Si un modelo ajusta bien los datos observados, entonces debe ocurrir también lo contrario: los datos generados a partir del modelo deben lucir similares a los datos observados. Es una verificación de consistencia interna. No todo modelo es perfecto, de tal manera que se considera útil un modelo que cumple con dicha consistencia interna con base en estadísticas de interés. Si interesaran la media y la dispersión, se chequea dicha consistencia. Puede que el valor mínimo no sea consistente, pero si no es de interés, el modelo podría seguir considerándose consistente para las necesidades para las que se genera.

La verificación se realiza gráfica y analíticamente.

Gráficamente, se genera un histograma de la estadística de interés, por ejemplo la mediana, por medio de un millar de muestras generadas a partir de la distribución marginal posterior y se compara con la estadística de interés T(y) calculada a partir de los datos observados. Tomando el ejemplo mencionado, se espera que la mediana de los datos observados esté dentro del intervalo que conforman los cuantiles 0.05 y 0.95 del histograma.

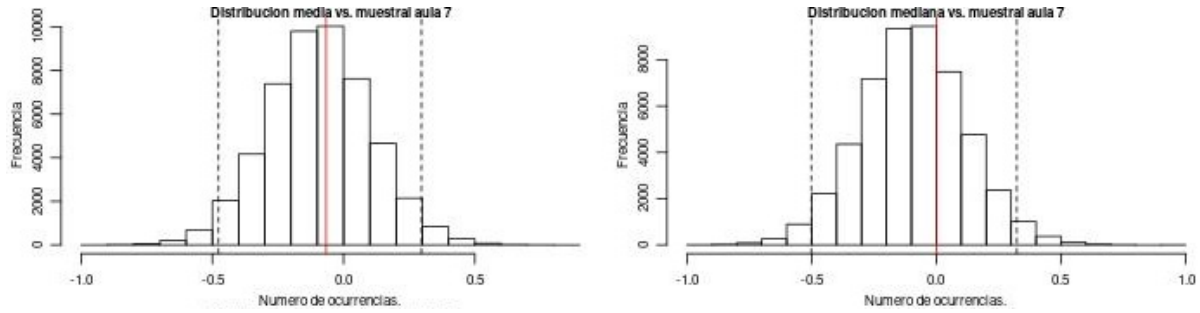



Figura 7: Comparación de la estadística de interés de los datos con la distribución de un muestreo a partir del modelo

El cálculo analítico es un equivalente. Se calcula el porcentaje de valores que superan al valor observado $P(T(y^{rep}) > T(y) \mid y)$. Se espera que dicho valor esté entre el 5 % y el 95 %. Al cálculo lo denominan *posterior predictive p-value (ppp)* por cuanto tiene alguna analogía con el valor p de la estadística frecuentista. Se diferencia en que la validación se usa para entender los límites de la aplicabilidad del modelo en la práctica, no para aceptar o rechazar la hipótesis nula[8] y que se interpreta directamente como porcentaje.

Para el presente trabajo se verificó media, mediana y desviación estándar.

El numeral 4.4 del libro de Hoff[11] expone cómo verificar el modelo predictivo posterior. También el numeral 24.1 de Gelman[6] expone los principios de la verificación del modelo predictivo posterior y su artículo[7] expone expresamente la idea.

### 4.4.3 Comparación de modelos: ECM y DIC

**Validación cruzada** Un modelo jerárquico representa bien la estructura del sistema educativo, pero son preferibles los modelos parsimoniosos, y tal vez no hace falta la representación jerárquica para modelar la evaluación. Para validar esto se comparan los modelos jerárquicos de dos y tres niveles frente al modelo base por medio del error cuadrático medio.

El proceso a realizar es generar una partición de la data con el objeto de realizar una validación cruzada k-fold. Si se selecciona un K = 10, entonces se realizan diez modelados, en cada uno asignándoles NA a los valores de uno de los subconjuntos. Asóciese con K el subíndice $k \in \{1, 2, 3, \ldots, 10\}$. Los NA son imputados $T$ veces, tantas como iteraciones realiza el MCMC (asóciese con T el subíndice $i \in \{1, 2, 3, \ldots, 10000\}$). Se obtienen así 10 matrices de dimensión $T * L$, donde $L$ es el número de estudiantes imputados.

Al calcular el promedio por columna, se obtiene un vector de tamaño $L$, con el valor medio de la imputación para cada estudiante ($\bar{y}_j$), donde j es el subíndice del estudiante.

A continuación se calcula el error cuadrático medio de la predicción: $\widehat{ECM}_k = \frac{\sum_{j=1}^{L}(y_j - \bar{y}_j)^2}{L}$, donde $y_j$ es el valor real del estudiante j.

Sobre los diez resultados se calcula la media del error cuadrático medio:

$$MECM = \frac{\sum_{k=1}^{10} \widehat{ECM}_k}{10}$$

Las matrices de dimensión $T * L$ también permiten calcular una imputación del error cuadrático dato por dato ($[y_j - \hat{y}_j]^2$), obteniendo para cada individuo una matriz de dimensión $T * 1$, que proporciona una distribución del error cuadrático por estudiante y por ende calcular un intervalo de confianza. Se denomina *comparación basada en el error de predicción*[8].

La validación cruzada permite aplicar paralelización del MCMC con el objeto de utilizar al 100 % de la capacidad del PC y reducir el tiempo de procesamiento.

Para evitar introducir variabilidad en la comparación del error cuadrático medio, la partición es la misma para los tres modelos.

Se espera que el modelo jerárquico de dos niveles, con cofactores, tenga un menor error cuadrático medio que el modelo jerárquico de tres niveles, y éste menor que el modelo base. De otro modo la complejidad del modelo debe justificarse en términos de análisis que no se puedan realizar con los más sencillos.

**DIC** Otra manera de comparar los modelos es mediante el DIC (Deviance Information Criterion). Al igual que sus equivalentes frecuentistas, su intención es castigar los modelos más complejos y verificar que son mejores a pesar de la complejidad introducida.

$$DIC = 2[-\ln \mathrm{P}(y \mid \hat{\theta}_{Bayes}) + \mathrm{P}_{\mathrm{DIC}}] \tag{43}$$

donde

$$\ln \mathrm{P}(y \mid \hat{\theta}_{Bayes}) = \sum_{k=1}^{M} \sum_{j=1}^{n_j} \sum_{i=1}^{n_{jk}} \ln\Big[\mathrm{N}(y_{ijk} \mid \bar{\theta}_{jk}, \bar{\sigma}_k^2)\Big]$$

y

$$\mathrm{P}_{\mathrm{DIC}} = 2\Big(\ln \mathrm{P}(y \mid \hat{\theta}_{Bayes}) - \frac{1}{S} \sum_{s=1}^{S} \log \mathrm{P}(y \mid \theta^{(s)})\Big)$$

$\mathrm{P}_{\mathrm{DIC}}$ es el número efectivo de parámetros.

A menor DIC, mejor el modelo.

El numeral 7.2 del libro de Gelman et al.[8] expone el Deviance Information Criterion y la validación cruzada como métodos de evaluación de modelos. Congdon también expone el tema en el capítulo 3.[3]. Estos autores desechan criterios como el Akaike Information Criterion (AIC) y el Bayes Information Criterion (BIC). AIC se desecha porque en los modelos de estructura jerárquica el número efectivo de parámetros, valor con que AIC *castiga*, dependen de la varianza de los parámetros del grupo-nivel (pag 172 de Gelman et al.). BIC lo desestiman porque no está diseñado para calcular el desempeño de un modelo a partir de una *muestra fuera del grupo* (pag 175 de Gelamn et al.). Otra medición que mencionan como alternativa y que sí estiman adecuada es el Watanabe-Akaike Information Criterion (WAIC). Consideran que es realmente Bayesiano, y que tiene una conexión más o menos directa con la validación cruzada (pag 182 de Gelman et al.).

# 5. Computación

## 5.1 Aspectos generales

Los modelos se implementaron en el paquete R (v 4.0.3), programados, sin hacer uso de paquetes para modelos Bayesianos. Los códigos se encuentran en Github y se puede dar acceso a ellos a petición.

A manera de referencia, el modelo jerárquico de tres niveles tomó, en un core intel i7, de 20 a 24 horas cada corrida. Y el modelo jerárquico de dos niveles de 10 a 12 horas. En un core intel i5 estos tiempos son de 30 a 36 horas y de 22 a 24 horas respectivamente.

Las cadenas resultantes de los samplers para cada índice ocupa un espacio en disco de 1.45 GB. 2.21 GB si se guardan las gráficas asociadas.

No se llevó a cabo traducción de códigos para que se ejecutase en lenguaje C++, lo cual haría los cálculos del orden de 100 veces más rápidos, por cuanto el tiempo disponible para la realización del proyecto no incluía el necesario para estudiar la implementación en C++ de modelos de datos complejos como los que requieren las distribuciones normales multivariadas. Se aplicó la paralelización para la realización de los cálculos de la validación cruzada.

## 5.2 Algoritmo

La especificación del modelo jerárquico de tres niveles implicó estimar 69 parámetros $\theta$, 28 parámetros: $\sigma^2$, $\mu$ y $\tau^2$; y un parámetro: $\alpha$, $\eta$, $\gamma$, $\kappa^2$, $\lambda$ y $\xi$. Un total de 159. (Ver Figura 3)

Para un conjunto de hiperparámetros dado: $\{a_0, b_0, c_0, d_0, e_0, f_0, g_0, h_0, u_0, v_0\}$, el algoritmo programado en R genera valores para la iteración $^{(b+1)}$ a partir de la iteración $^{(b)}$ de la siguiente manera:

1. Muestrear cada $\theta_{jk}^{(b+1)}$ de una distribución $N(\mu_k^{(b)}, \tau_k^{2(b)})$
2. Muestrear cada $\sigma_k^{2(b+1)}$ de una distribución $IGamma(\alpha^{(b)}, \eta^{(b)})$
3. Muestrear cada ${\mu_k}^{(b+1)}$ de una distribución $N(\gamma^{(b)}, {\kappa^2}^{(b)})$
4. Muestrear cada $\tau_k^{2(b+1)}$ de una distribución $IGamma(\lambda^{(b)}, \xi^{(b)})$
5. Muestrear $\alpha^{(b+1)}$ de un vector de valores $exp(valor - max(valor))$, donde $valor = 28n\log(\eta^{(b)}) - (n+1)\sum_{k=1}^{28}\log(\sigma_k^{2(b)}) - 28\Gamma(n) - na_0$, con $n \in \{1, 2, 3, \ldots, 1000\}$. El anexo D presenta los detalles que justifican esta formulación.
6. Muestrear $\eta^{(b+1)}$ de una distribución $Gamma(b_0, c_0)$
7. Muestrear $\gamma^{(b+1)}$ de una distribución $N(g_0, h_0^2)$
8. Muestrear ${\kappa^2}^{(b+1)}$ de una distribución $IGamma(u_0, v_0)$
9. Muestrear $\lambda^{(b+1)}$ de un vector de valores $exp(valor - max(valor))$, donde $valor = 28n\log(\xi^{(b)}) - (n+1)\sum_{k=1}^{28}\log(\tau_k^{2(b)}) - 28\Gamma(n) - nd_0$, con $n \in \{1, 2, 3, \ldots, 1000\}$. Ver anexo D.
10. Muestrear $\xi^{(b+1)}$ de una distribución $Gamma(e_0, f_0)$

donde $k$ representa un establecimiento educativo (en total son veintiocho) y $jk$ al aula j del establecimiento educativo k (en total son 69).

## 5.3 Elicitación de hiperparámetros

Una vez programados los samplers del modelo jerárquicos de tres niveles, se plantea ponerlos a prueba utilizando datos simulados. Con el objeto de verificar que la simulación se está comportando bien se seleccionan hiperparámetros que generen varianzas muy informativas, especificando un coeficiente de variación pequeño, por ejemplo, 0.05. En cambio, cuando se modela, se seleccionan hiperparámetros de tal modo que la distribución no sea informativa, con varianzas amplias de acuerdo a lo aconsejado por Gelman et al.[8].

### 5.3.1 Simulación de datos

Para la simulación de datos del modelo jerárquicos de tres niveles basta con representar los datos del primer nivel desde el segundo nivel (Ver Figura 3).

Se especifica $\alpha$ y $\eta$, parámetros de forma y escala respectivos de $\sigma^2 \sim Gamma(\alpha, \eta)$.

$\alpha \propto e^{-a_0\alpha}$ con $\alpha \in \mathbb{N}$. $CV_{\sigma^2} = \frac{1}{\sqrt{(\alpha-2)}}$, por ser gamma inversa, así que $E(\alpha) = CV_{\sigma^2}^{-2} + 2$. Si $CV_{\sigma^2} = 0{,}05$, entonces $E(\alpha) = 402$ y $a_0 = 1/E(\alpha) = 0.0025$.

Para que la esperanza de $\sigma^2$ sea pequeña y $\alpha > 1$, los parámetros $b_0$ y $c_0$ deben tener en cuenta que[3] $E(\sigma^2) = \frac{\eta}{E(\alpha)-1}$, y $E(\eta) = b_0 * c_0$. Sea fijada $E(\sigma^2) = 0{,}005$, entonces $E(\eta) = 0{,}005 * 401 \sim 2$. Si $b_0 > 1$ y $b_0 * c_0 = 2$, implica que $b_0 = 2/c_0 > 1$, por tanto $c_0 < 2$. Se escogió $b_0 = 400$ y $c_0 = 0.005$. Implica $CV\eta = 0.05$.

Todo lo dicho para $\alpha$ y $\eta$ aplica para $\lambda$ y $\xi$ respectivamente.

$\gamma \sim N(g_0, h_0^2)$. Sea $g_0 = 0.1$, y $h_0 = 0.05$. Como se simula sólo desde el nivel 2, $\gamma$ se fija en la simulación a los valores de la esperanza, es decir, $\gamma = 0.1$.

$\kappa^2 \sim IGamma(u_0, v_0')$. Sea $u_0 = 402$ y $v_0' = 0.05$ (ver Figura 9). Como $E(\kappa^2) = \frac{v_0'}{u_0-1}$, se deja $\kappa^2$ fija con valor 0.000125.

Se puede especificar ahora $\mu_k$ y $\tau_k^2$, media y varianza respectiva de $\theta_{jk}$.

| **Parámetro** | **Hiperparámetro en modelado** | **Hiperparámetro en simulación** | **Valor esperado parámetro en simulación** |
|---|---|---|---|
| $\alpha \sim \exp(a_0)$ | $a_0 = 0{,}01$ | $a_0 = 0{,}1$ | 10 |
| $\eta \sim Gamma(b_0, c_0)$ | $b_0 = 1, c_0 = 0{,}01$ | $b_0 = 4, c_0 = 2$ | 2 |
| $\gamma \sim N(g_0, h_0^2)$ | $g_0 = 0, h_0^2 = 1$ | $g_0 = 0{,}1,\ h_0^2 = 0{,}0025$ | 0.1 |
| $\kappa^2 \sim IGamma(u_0, v_0)$ | $u_0 = 3, v_0 = 0{,}2$ | $u_0 = 402, v_0 = 0{,}05$ | $1{,}25 * 10^{-4}$ |
| $\lambda \sim \exp(d_0)$ | $d_0 = 0{,}01$ | $d_0 = 0{,}1$ | 10 |
| $\xi \sim Gamma(e_0, f_0)$ | $e_0 = 1, f_0 = 0{,}01$ | $e_0 = 4, f_0 = 2$ | 2 |

Cuadro 4: Resumen hiperparámetros

[3]Se utiliza $E(\alpha)$ en vez de $\alpha$ por cuanto en la simulación no se toma $\alpha$ como una variable aleatoria sino como un valor.

### 5.3.2 Modelado de datos

Para el modelado de datos deben especificarse los hiperparámetros de $\alpha$, $\eta$, $\gamma$, $\kappa^2$, $\lambda$ y $\xi$.

Por especificación se modela $\gamma \sim N(g_0, h_0)$. Se selecciona una media, por ejemplo, ($g_0 = 0$) para que el valor esperado no tenga un valor a priori. Y una desviación estándar amplia. Una desviación estándar de la diferencia de medias en este tipo de programas de educación física, recreación y deporte de magnitud 1 es amplia. Sobre una N(0,1), el percentil 1 y 99 es ±2,326. Es decir, es válido que $h_0 = 1$.

$\kappa^2 \sim IGamma(u_0, v_0)$. El parámetro de forma, $u_0$, controla la altura. A mayor $u_0$, mayor la altura de la función de densidad. El parámento de escala, $v_0$, controla la dispersión. A mayor $v_0$, mayor dispersión. Pero el modelamiento en el presente trabajo se está realizando respecto a la rata: $\kappa^2 \sim IGamma(u_0, v_0')$, con $v_0'$ como inverso de $v_0$, la cual representa la precisión, por tanto, a mayor $v_0'$, mayor precisión.

Entonces $E(\kappa^2) = \frac{v_0'}{(u_0-1)}$, con $u_0 > 1$, y $Var(\kappa^2) = \frac{v_0'^2}{(u_0-1)^2(u_0-2)}$ con $u_0 > 2$, por consiguiente, $CV_{\kappa^2} = \frac{\sqrt{Var(\kappa^2)}}{E(\kappa^2)} = \frac{\frac{v_0'}{(u_0-1)(u_0-2)^{1/2}}}{\frac{v_0'}{(u_0-1)}} = \frac{v_0'(u_0-1)}{v_0'(u_0-1)(u_0-2)^{1/2}} = (u_0-2)^{-1/2}$. Obsérvese que el $CV_{\kappa^2}$ es el mismo bajo parametrización de escala o de rata. Despejando, $u_0 = CV_{\kappa^2}^{-2} + 2$. Luego, despejando $v_0'$ de la esperanza: $v_0' = E(\kappa^2)(u_0 - 1)$.

Sea un coeficiente de variación alto: $CV_{\kappa^2} = 1$, entonces $u_0 = 3$, y sea $E_0(\kappa^2) = 0.25$, por tanto $v_0'$ = 0.5. El bajo valor de $v_0'$ implica baja precisión.

$\alpha \propto exp(-\alpha a_0)$. En la distribución exponencial el CV siempre es 1. $Var(\alpha) = 1/a_0^2 = 640000$, entonces $a_0 = 0.00125$. Implica que $E(\alpha) = 800$.

$\eta \sim Gamma(b_0, c_0)$, $c_0$ como rata. Con un coeficiente de variación grande: $CV_{\eta_{rata}} = 1 = b_0^{-1/2}$ obtendríamos un $b_0 = CV_\eta^{-2} = 1$. Y sea $E(\eta) = 10 = \frac{b_0}{c_0}$, entonces, $c_0 = b_0/E(\eta) = 0.1$.

$\lambda$ se modela exactamente igual que $\alpha$, por tanto sea $d_0 = 0.00125$.

$\xi \sim Gamma(e_0, f_0)$ del mismo modo que $\eta$. Si $CV_\xi = 1$, entonces $f_0 = 0.1$. Y sea $E(\xi) = 10$, entonces, $e_0 = f_0 * E(\xi) = 1$.

Para los modelos jerárquicos de dos niveles los hiperparámetros $a_0$, $b_0$ y $c_0$ son los mismos que para el modelo jerárquico de tres niveles.

En relación a los hiperparámetros de $\Psi$, se busca que conformen una distribución no informativa. La media se centra en cero:

$$\mu^T = [0, 0, 0]$$

La covarianza se asigna en cero para hacerla no informativa. En cambio, la varianza de asigna amplia con el mismo objetivo. Entonces, la matriz de varianza-covarianza que se asigna es:

$$\mathbf{\Lambda} = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{bmatrix}$$

En relación a los hiperparámetros de la distribución de $\Sigma$, Hoff [11] aconseja una matriz $S_0$ igual a $\Lambda$, pero centrada libremente alrededor de $p + 2$, es decir, $\nu_0 = 3$ ó $\nu_0 = 5$.

$$\mathbf{S}_0 = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{bmatrix}$$

La matriz de parámetros Beta se inició con un vector centrado en cero.

### 5.3.3 Gráficas

Se presentan gráfica ilustrativas del planteamiento dado en los dos literales anteriores. Las informativas corresponden a la simulación de datos. Las no informativas, al modelado.

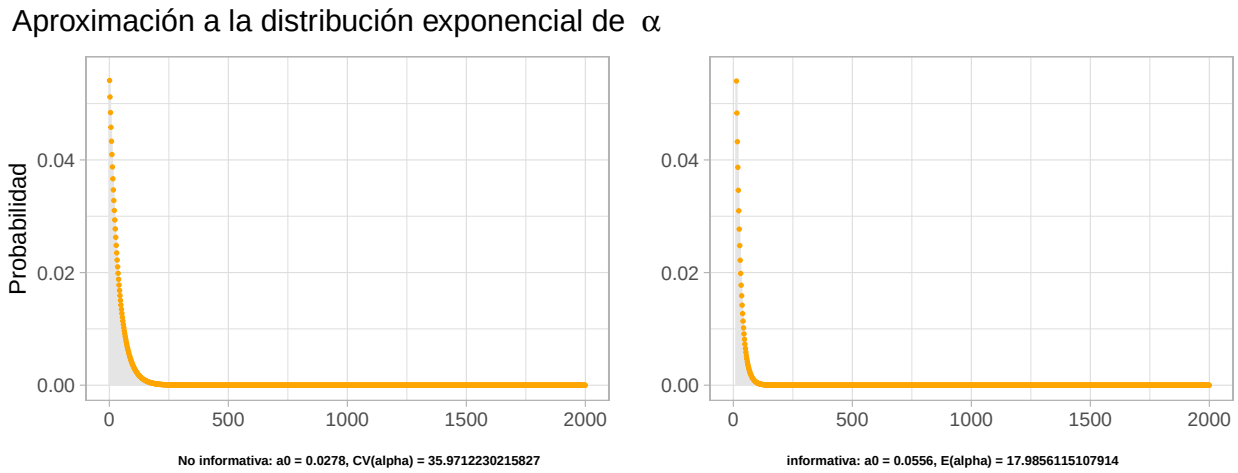
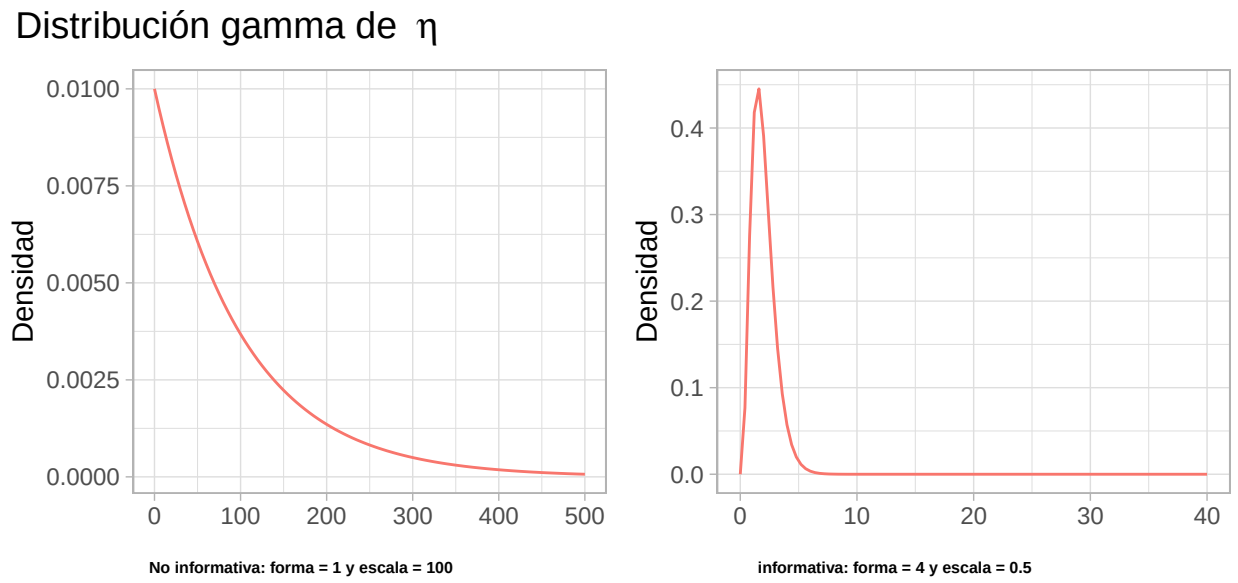



Figura 8: Elicitación de parámetros para las distribuciones de gamma y alpha



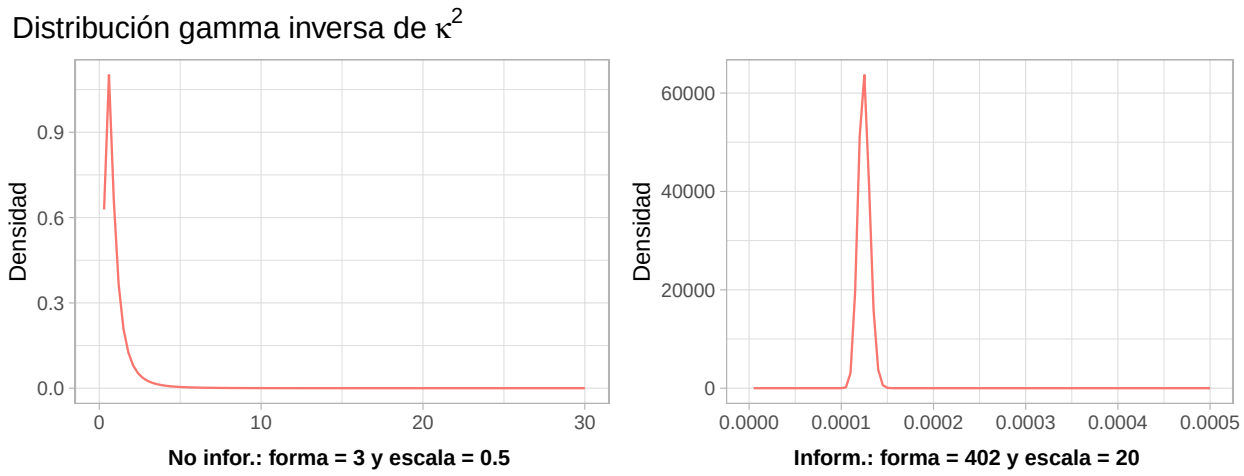



Figura 9: Elicitación de parámetros para las distribuciones de eta y kappa

## 5.4 Validación de los modelos

Se procede a aplicar la metodología de validación de los modelos para determinar si es dable llegar a conclusiones acerca del problema planteado.

Se realiza el proceso completo para el indice de Liderazgo efectivo, modelo por modelo. Se presenta en Anexo F la validación de los modelos sobre los otros dos índices por cuanto los resultados son en todo similares.

### 5.4.1 Modelo base

El cálculo de la Log-verosimilitud del MCMC para el modelo base es:

$$logP\left(\tilde{y} \mid \theta^{(s)}, (\sigma^2)^{(s)}\right) = \Sigma_{k=1}^{M}\Sigma_{j=1}^{n_k}\Sigma_{i=1}^{n_{jk}} N(\theta_k, \sigma^2)$$

Dónde el superíndice (s) representa la iteración que corresponde.

$$s \in \{1, 2, 3, \ldots, S\}$$

S es el número total de iteraciones.

La gráfica de la Figura 10 presenta la convergencia del modelamiento para $logP\left(\tilde{y} \mid \theta^{(s)}, (\sigma^2)^{(s)}\right)$.

El modelo base es sencillo. Se realizaron 52,000 iteraciones, con un calentamiento de la serie de 2,000 iteraciones y guardando los resultados cada 5 iteraciones con el objeto de obtener finalmente 10,000 muestras. Se observa un muy buen comportamiento, con nula auto-regresión entre las muestras. Bajo pruebas formales, realizadas con el paquete *coda* de R, la cadena converge.

El tamaño efectivo de $\theta$ y $\sigma^2$ es 10,050 y 10,000 respectivamente.

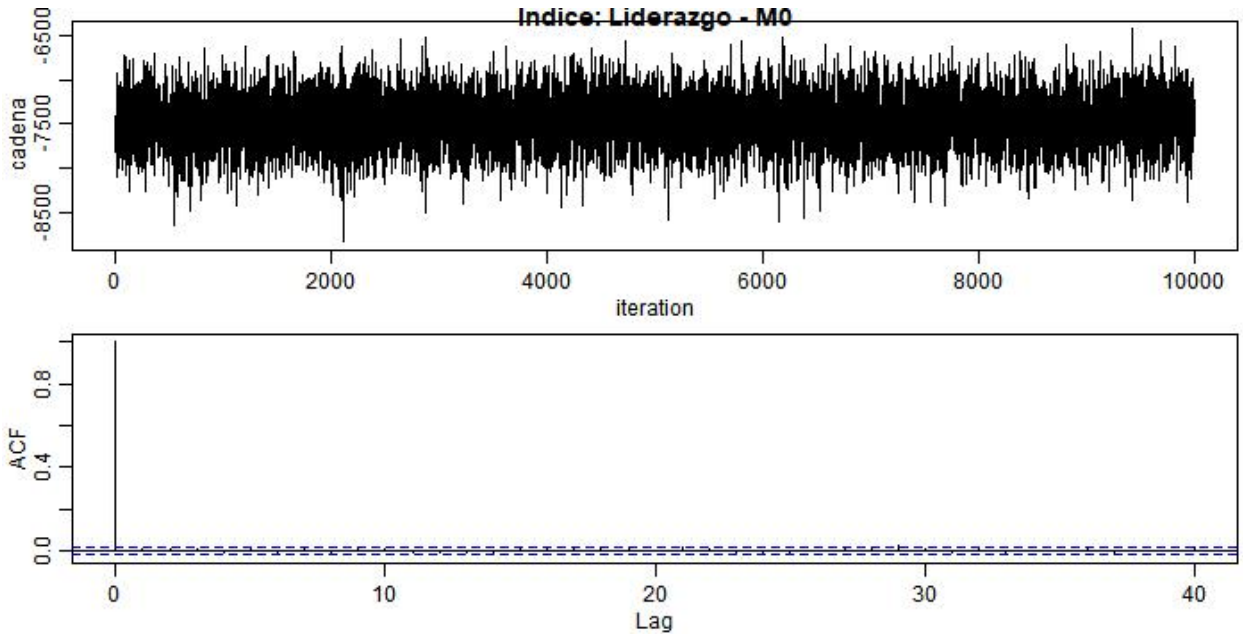



Figura 10: Convergencia del modelamiento para logP

Cuadro 5: Errores estándar del modelo base

| Distribución | Error estándar |
|---|---|
| Theta | 0.0003655 |
| Sigma2 | 0.0000323 |

Los errores estándar son adecuados.

La validación cruzada realizada sobre el modelo base obtuvo un error cuadrático medio (AMSE) de 1.102.

Se presenta a continuación los ajustes asociados al Deviance Information Criterion del Modelo base:

$$DIC = 2[-\ln \mathrm{P}(y \mid \hat{\theta}_{Bayes}) + \mathrm{P}_{\mathrm{DIC}}]$$

donde

$$\ln \mathrm{P}(y \mid \hat{\theta}_{Bayes}) = \sum_{k=1}^{M} \sum_{j=1}^{n_j} \sum_{i=1}^{n_{jk}} \ln \Big[ \mathrm{N}(y_{ijk} \mid \bar{\theta}_k, \bar{\sigma}^2) \Big]$$

y

$$\mathrm{P}_{\mathrm{DIC}} = 2\Big( \ln \mathrm{P}(y \mid \hat{\theta}_{Bayes}) - \frac{1}{S} \sum_{s=1}^{S} \log \mathrm{P}(y \mid \theta^{(s)}) \Big)$$

El DIC del modelos es -944,313.89.

Finalmente, se verifica la consistencia interna del modelo. Como se enunció en la metodología, se utilizaron tres estadísticas para determinar la coherencia interna: media, mediana y desviación estándar:

Las estadísticas escogidas son consistentes para algunos establecimientos educativos, para otras no.

La manera adecuada de evaluarlo es determinar el *predictive posterior p-value (ppp)*. Es de esperar que dicho valor se sitúe entre el 2.5 % y el 97.5 % para que sea consistente.

Cuadro 6: Porcentaje de establecimientos educativos, en el modelo base, con un predictive posterior p-value dentro de los rangos esperados

| | Media | Mediana | Desviación estándar |
|---|---|---|---|
| Establecimiento educativo | 100 | 60.7 | 0 |

El cuadro 6 presenta el porcentaje de establecimientos educativos cuyo ppp-value está dentro de los límites especificados. Se observa que el modelo tiene problemas en representar bien la desviación estándar. Y el valor de la mediana también es bajo. Esto todavía permite responder las preguntas que se plantearon al inicio del trabajo por cuanto se refieren todas a la media.

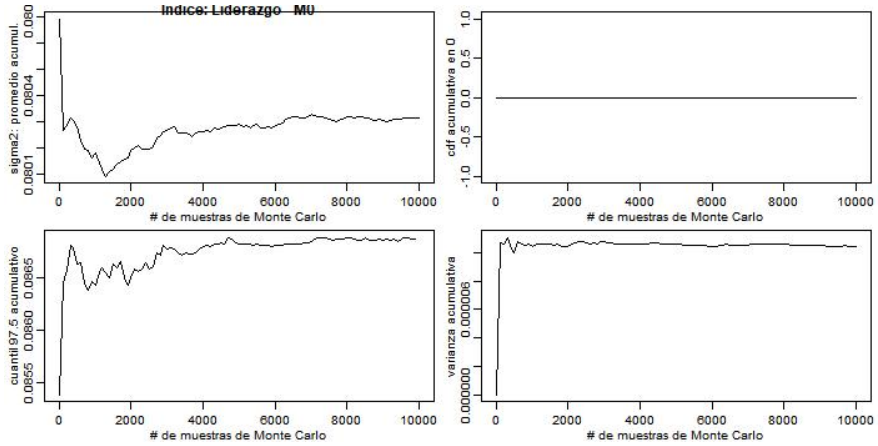
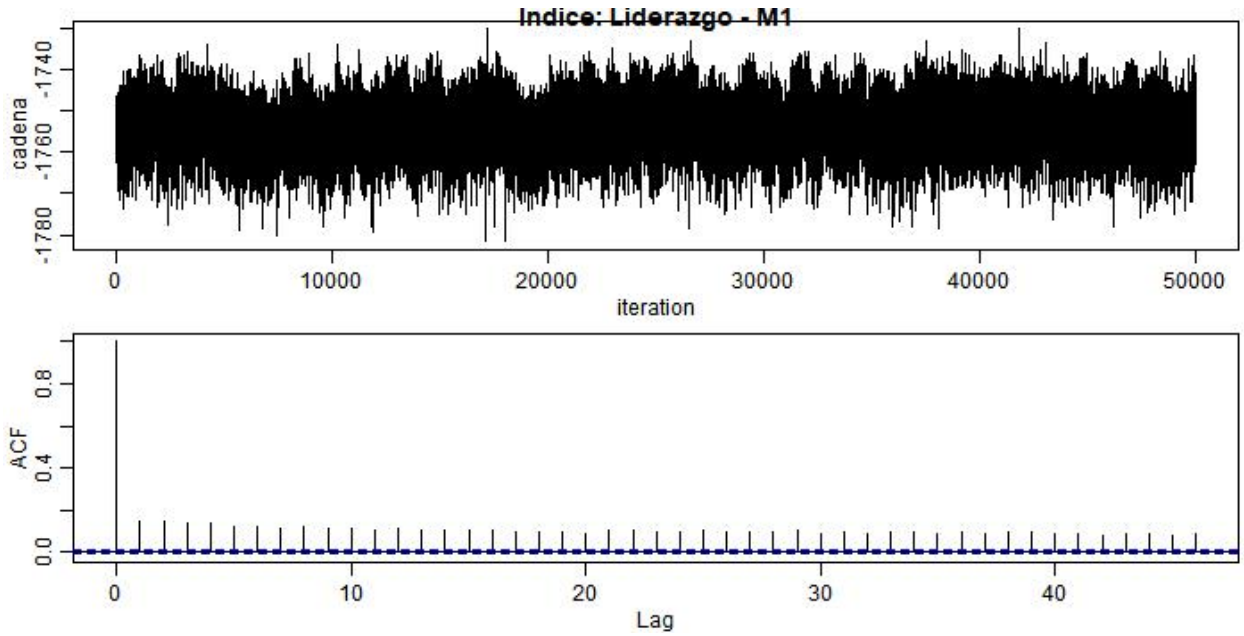



Figura 11: Convergencia del parámetro sigma2

La Figura 11 presenta que se llegó a la convergencia en el parámetro $\sigma^2$ del modelo base, en consecuencia, no es problema atribuible a un bajo número de iteraciones, sino a que la especificación no logra representar adecuadamente los datos.

El panel superior izquierdo de la Figura 11 presenta la media acumulada del parámetro. El panel superior derecho presenta el porcentaje acumulado de datos menores que cero. El panel inferior izquierdo presenta el acumulado del valor del percentil $p_{97,5}$ El panel inferior derecho presenta la varianza acumulada del parámetro.

### 5.4.2 Segundo modelo: jerárquico de tres niveles




La convergencia es adecuada, pero presenta algún nivel de autocorrelación. Bajo pruebas formales, realizadas con el paquete *coda* de R, la cadena converge.

Obsérvese que el número de iteraciones fue de 50,000 para el modelo. Esto merece una explicación.

El modelo jerárquico de tres niveles es el más complejo. Con 52,000 iteraciones se obtenía una alta correlación en los parámetros del nivel 3, así que se optó por correr 1,020,000 iteraciones, con un calentamiento de la serie de 20,000 iteraciones y guardando los resultados cada 20 iteraciones con el objeto de obtener finalmente 50,000 muestras. ¿Por qué más muestras? Para obtener tamaños de muestra efectivas suficientes para compararse con el modelo base. El comportamiento de las variables de interés: $\theta_{jk}$ y $\sigma_k^2$ es bueno, con muestras efectivas de 22,395 y 38,101. También los tamaños de muestra efectiva de $\mu_k$, $\gamma$ y $\kappa^2$ fueron buenos: 39,900, 44,813 y 47,501 respectivamente. Corresponde al modelamiento de la media de $\theta$. Pero los tamaños de muestra efectiva de $\tau^2$, $\lambda$ y $\xi$ fueron pésimos: 142, 134 y 78 respectivamente. Corresponden al modelamiento de la varianza de $\theta$.

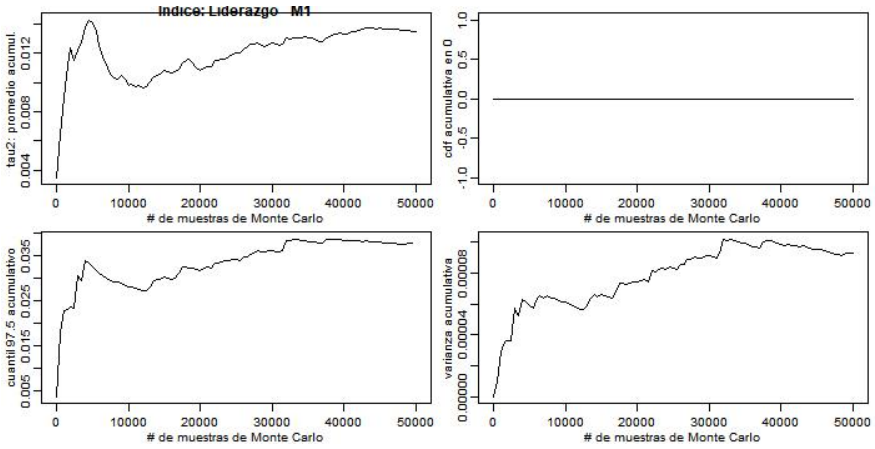



Figura 12: Convergencia del parámetro tau2

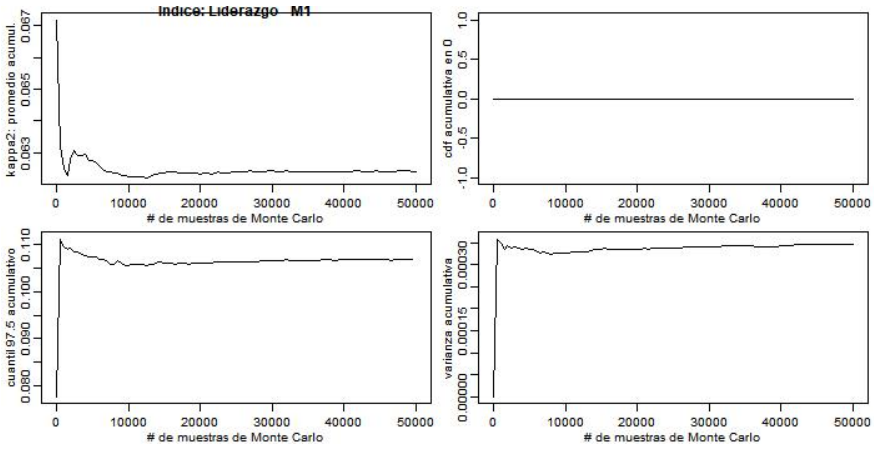



Figura 13: Convergencia del parámetro kappa2

En la Figura 12 se observa cómo no se logra estabilización (convergencia) del parámetro $\tau^2$, del nivel 2. Compárese con el comportamiento del parámetro $\kappa^2$ (Figura 13) del nivel 3, el cual sí presenta estabilización.

Sujeto a que no se lograrían tamaños de muestra efectivos adecuados para los parámetros $\tau^2$, $\lambda$ y $\xi$, los siguientes índices sociogrupales se calcularon con 25,000 iteraciones, suficientes para alcanzar un tamaño efectivo de 10,000 en los restantes parámetros.

Cuadro 7: Errores estándar del modelo jerárquico de tres niveles

| Distribución | Error estándar |
|---|---|
| Theta | 0.0009233 |
| Sigma2 | 0.0007879 |
| Mu | 0.0006638 |
| Tau2 | 0.0008072 |
| Alpha | 0.1564784 |
| Eta | 0.1410855 |
| Gamma | 0.0002666 |
| Kappa2 | 0.0000853 |
| Lambda | 22.1563488 |
| Xi | 0.5945697 |

Se observa un error estándar aparentemente alto para $\alpha$, $\eta$, $\lambda$ y $\xi$. No obstante, en relación a las medias no son valores altos (la media entre el EE es, respectivamente, 137.2, 133.6, 23.3, 10.5).

El modelo jerárquico de tres niveles obtuvo un AMSE de 1.088. Una desmejora del 1.3 % respecto al modelo base.

El DIC del modelo es -26,489.76.

El modelo jerárquico de tres niveles obtiene un desempeño inferior, medido mediante el DIC, al modelo base. No obstante, es consistente para las tres estadísticas tanto a nivel de establecimiento educativo, como de aula.

Cuadro 8: Porcentaje de establecimientos educativos, en el modelo jerárquico de tres niveles, con un predictive posterior p-value dentro de los rangos esperados

| | Media | Mediana | Desviación estándar |
|---|---|---|---|
| Aula | 100 | 100 | 92.8 |
| Establecimiento educativo | 100 | 100 | 100.0 |

### 5.4.3 Segundo modelo: jerárquico de dos niveles, con covariables

El cálculo de la Log-verosimilitud del MCMC para los modelos jerárquicos de dos niveles es:

**A nivel de aula**

$$logP\left(\tilde{y}_{ijk} \mid \boldsymbol{\beta}_{jk}^{(s)}, (\sigma_k^2)^{(s)}\right) = \Sigma_{k=1}^{M}\Sigma_{j=1}^{n_k}\Sigma_{i=1}^{n_{jk}} N(\boldsymbol{\beta}_{jk}^{(s)} X_{ijk}^T, (\sigma_k^2)^{(s)})$$

**A nivel de establecimiento educativo**

$$logP\left(\tilde{y}_{ijk} \mid \boldsymbol{\beta}_{k}^{(s)}, \sigma_k^{2(s)}\right) = \Sigma_{k=1}^{M}\Sigma_{j=1}^{n_k}\Sigma_{i=1}^{n_{jk}} N(\boldsymbol{\beta}_{k}{}^{(s)} \boldsymbol{X}_{ijk}^T, \sigma_k^{2(s)})$$

Dónde el superíndice (s) representa la iteración que corresponde.

$$(s) \in \{1, 2, 3, \ldots, S\}$$

S es el número total de iteraciones.

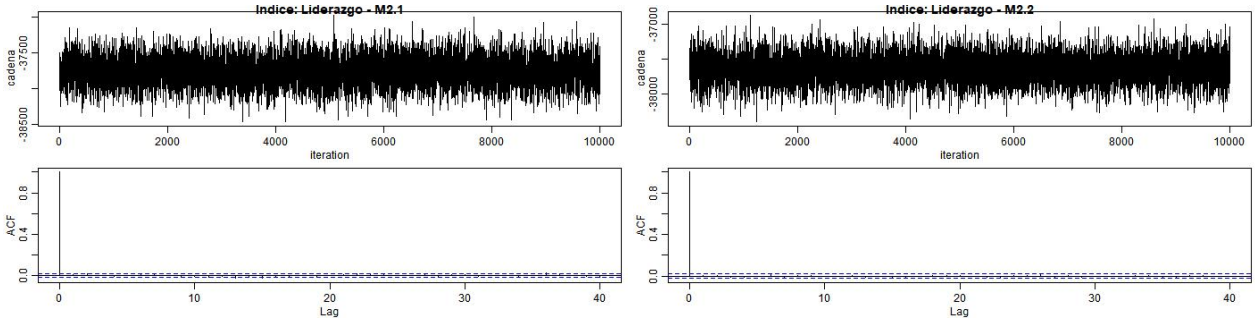



Figura 14: Convergencia de los modelos para el indice Liderazgo

Las gráficas de la Figura 14 presentan la cadena de la logverosimilitud para cada uno de los modelos. Las dos opciones del modelo convergieron de manera adecuada. Para ambos se utilizó una configuración de iteraciones de MCMC: 102,500 iteraciones, con un calentamiento de la serie de 2,500 iteraciones, guardada de los resultados cada 10 iteraciones con el objeto de obtener 10,000 muestras.

Cuadro 9: Errores estándar del modelo M2.1 para el quinto establecimiento educativo

| Distribución | Error estándar |
|---|---|
| Beta_0 | 0.0004631 |
| Beta_1 | 0.0006002 |
| Sigma2 | 0.0000057 |

Los tamaños efectivos para, por ejemplo, $\beta_0$, $\beta_1$ y $\sigma^2$ del primer modelo jerárquico de dos niveles fueron: 9,971, 10,040 y 9,983 respectivamente.

Cuadro 10: Errores estándar del modelo M2.2 para el quinto establecimiento educativo

| Distribución | Error estándar |
|---|---|
| Beta_0 | 0.0213792 |
| Beta_1 | 0.0214437 |
| Beta_2 | 0.0376082 |
| Beta_3 | 0.0216032 |
| Sigma2 | 0.0000055 |

Los tamaños efectivos para el segundo modelo jerárquico de dos niveles fueron: 9,916, 9,857, 9,693, 9,709 y 1,714,884 respectivamente.

Obsérvese que para el parámetro $\sigma^2$ el *tamaño de muestra efectivo* es muy grande. Esto ocurre cuando hay autocorrelaciones negativas en las MCMC. El tamaño de muestra efectivo se utiliza como una aproximación al tamaño de las cadenas si se hubieran obtenido muestras independientes. Cuando hay una correlación negativa la varianza del estimador de las MCMC correlacionadas puede ser menor que la varianza del estimador de las muestras independientes, lo que conduce a un tamaño de muestra efectivo más grande. No es problema.

En comparación con el modelo por sexo, el modelo por grado obtiene un mayor error estándar en sus coeficientes $\boldsymbol{\beta}$ (la media de los betas entre sus respectivos EE es 16.2, 41.5, 0.8, 43.7).

El primer modelo jerárquico de dos niveles obtuvo un AMSE de 1.154 en la validación cruzada. Una mejora del 4.7 % respecto al Modelo base y del 6.1 % respecto al modelo jerárquico de tres niveles.

Se presenta a continuación el ajuste asociado al Deviance Information Criterion de los modelos jerárquicos de dos niveles:

**A nivel de aula**

$$DIC = 2[-\ln \mathrm{P}(y \mid \hat{\beta}_{Bayes}) + \mathrm{P}_{\mathrm{DIC}}]$$

donde

$$\ln \mathrm{P}(y \mid \hat{\beta}_{Bayes}) = \sum_{k=1}^{M} \sum_{j=1}^{n_j} \sum_{i=1}^{n_{jk}} \ln \Big[ \mathrm{N}(y_{ijk} \mid X_{ijk}^T \bar{\boldsymbol{\beta}}_{jk}, \bar{\sigma}_k^2) \Big]$$

y

$$\mathrm{P_{DIC}} = 2\Big(\ln \mathrm{P}(y \mid \hat{\beta}_{Bayes}) - \frac{1}{S}\sum_{s=1}^{S} \log \mathrm{P}(y \mid X_{ijk}^T \boldsymbol{\beta}_{jk}^{(s)})\Big)$$

**A nivel de establecimiento educativo**

$$DIC = 2[-\ln \mathrm{P}(y \mid \hat{\beta}_{Bayes}) + \mathrm{P_{DIC}}]$$

donde

$$\ln \mathrm{P}(y \mid \hat{\beta}_{Bayes}) = \sum_{k=1}^{M}\sum_{j=1}^{n_j}\sum_{i=1}^{n_{jk}} \ln\Big[\mathrm{N}(y_{ijk} \mid X_{ijk}^T \bar{\boldsymbol{\beta}}_k, \bar{\sigma}_k^2)\Big]$$

y

$$\mathrm{P_{DIC}} = 2\Big(\ln \mathrm{P}(y \mid \hat{\beta}_{Bayes}) - \frac{1}{S}\sum_{s=1}^{S} \log \mathrm{P}(y \mid X_{ijk}^T \boldsymbol{\beta}_{k}^{(s)})\Big)$$

El DIC de los modelos es:

Cuadro 11: Deviance Information Criterion por modelo

| Modelo | DIC |
|---|---|
| Sexo | 56,477 |
| Grado | 83,633 |

Los modelos jerárquicos de dos niveles obtienen un DIC muy superior en valor, muy inferior en desempeño al modelo base.

A pesar del DIC mayor, los modelos jerárquicos de dos niveles permiten realizar inferencias sobre sexo y grado respectivamente, lo cual es una ganancia frente al aparente mejor desempeño de los otros dos modelos.

Cuadro 12: Porcentaje de aulas, en los modelos jerárquicos de dos niveles, con un predictive posterior p-value dentro de los rangos esperados

| | Media | Mediana | Desviación estándar |
|---|---|---|---|
| Sexo a nivel de Aula | 100 | 42.0 | 0 |
| Grado a nivel de establecimiento educativo | 100 | 46.4 | 0 |

Los modelos jerárquicos de dos niveles tienen un desempeño regular en la representación de lo que ocurre en las aulas o establecimientos educativos para las estadísticas de mediana y desviación estándar, sobre todo a nivel de desviación estándar. Pero para la media la representación es adecuada.

Las figuras 15 y 16 presentan que también los modelos jerárquicos de dos niveles logran convergencia del parámetro $\sigma^2$.

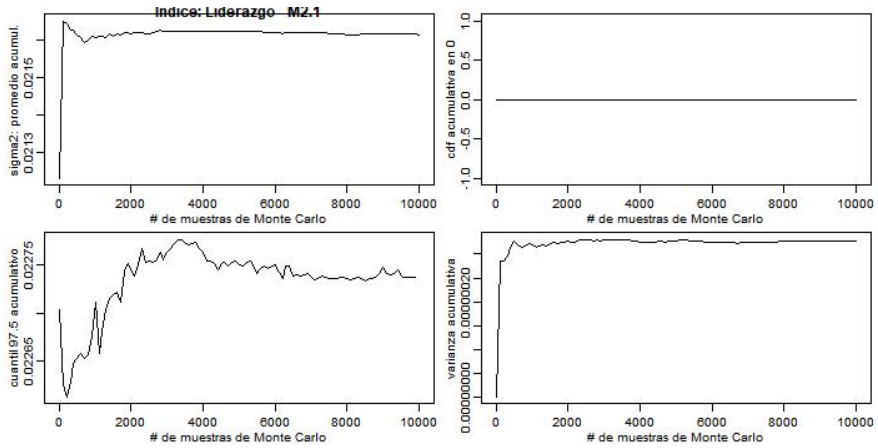



Figura 15: Convergencia del parámetro sigma2 para el primer modelo jerárquico de dos niveles

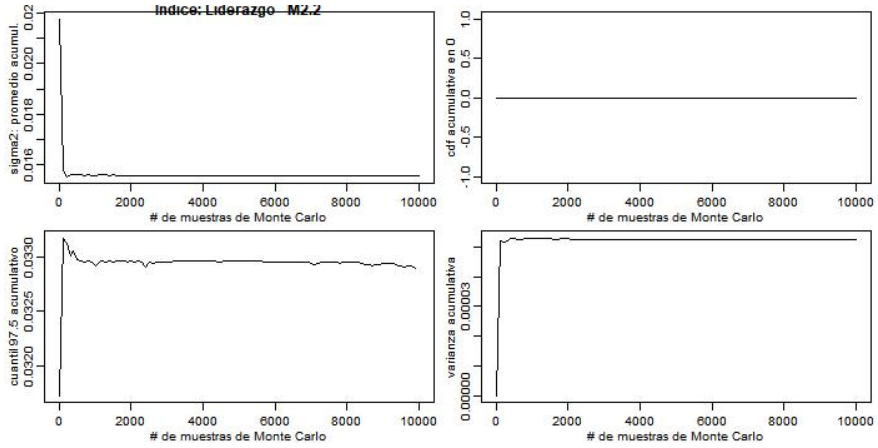



Figura 16: Convergencia del parámetro sigma2 para el segundo modelo jerárquico de dos niveles

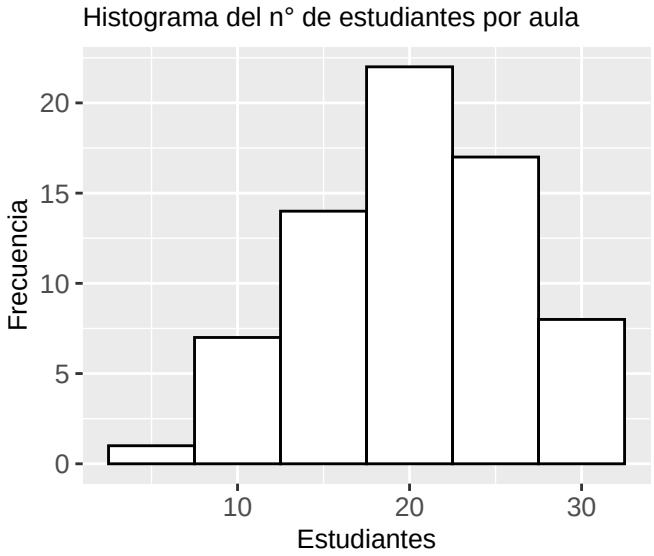
# 6. Discusión de los resultados

La aplicación práctica del modelamiento, explicada en la Sección de planteamiento del problema, se realiza sobre tres indices medidos en el año 2018 para veintiocho instituciones educativas. En cada una participan entre 1 y 5 aulas, la mayoría con 2 ó 3 aulas, para un total de sesenta y nueve aulas, y más de mil trescientos estudiantes.

Cuadro 13: Tabla de frecuencia de número de establecimientos educativos según número de aulas

| N° de aulas | N° de establecimientos educativos |
|---|---|
| 1 | 2 |
| 2 | 14 |
| 3 | 10 |
| 4 | 1 |
| 5 | 1 |

El número de estudiantes evaluados por aula varía entre 7 y 32 estudiantes.




Los indices analizados son *Liderazgo efectivo*, *Habilidades comunicativas* y *Respeto hacia los demás y hacia sí mismo*. Las tres son habilidades socio-grupales.

Liderazgo efectivo entendido como la capacidad para ejercer liderazgo en la ejecución de una tarea.

Habilidades comunicativas definido como logro de una comunicación efectiva con pares y docentes, en contextos interpersonales y en grupo.

Respeto hacia los demás y hacia sí mismo en términos de que el joven muestra claridad acerca del comportamiento adecuado frente a los límites ligados al espacio personal, cuerpo y tiempo propios y de los demás. Presenta capacidad para respetarlos y hacerlos respetar de forma adecuada.

Las preguntas que los financiadores y operadores de la intervención se plantearon son cuatro:

1. ¿Cuál es la probabilidad de que el valor de la diferencia entre la medición de los índices diferencia analizados sea mayor a cero? Les interesa de manera desagregada.

2. ¿La intervención logra efectos uniformes en todas las instituciones educativas?
3. ¿Hay brecha entre los sexos? ¿Es uniforme dicha brecha?
4. ¿La intervención obtiene resultados más convenientes en unos grados que en otros?

Para responder estas preguntas se construyeron tres modelos. El primer modelo es muy sencillo y se usa para comparar los dos restantes. El segundo y el tercero son modelos jerárquicos cuya formulación se expone en la Sección 4. E segundo es de tres niveles (Ver Figura 3) y el tercero de dos niveles (Ver figuras 4 y 5). El tercer modelo tiene dos formulaciones ligeramente diferentes. Se presenta por aula o por establecimiento educativo. La Sección 4 ilustra cómo es posible hacer una pequeña variación para cambiar el nivel de detalle. Este tercer par de modelos introduce variables indicadoras o dummies para evaluar el efecto por sexo o por grado, respectivamente.

El proceso de validación mostró que ambos modelos jerárquicos son aptos para modelar la media de los índices diferencia, por tanto, son convenientes para responder las preguntas. El primer modelo jerárquico permite contestar las dos primeras preguntas. El segundo modelo jerárquico permite contestar las dos últimas. Se analiza la respuesta por sexo a nivel de aula, ya que los establecimientos educativos manejan grupos mixtos. Y se analiza la respuesta por grado a nivel de establecimiento educativo, ya que no tiene sentido realizarlo a nivel de aula.

Se procede a responder las preguntas por cada indice.

### 6.1 Liderazgo efectivo

Como ya se mencionó, la primera y segunda pregunta se responden con el *modelo jerarquíco de tres niveles*.

El gráfico de oruga de la Figura 17 identifica el valor medio y el intervalo de confianza del índice diferencia. Donde la intervención es significativa y positiva se presenta el intervalo de confianza en color verde. Si fuese significativa y negativa se presentaría de color rojo. La intervención fue significativa y positiva para seis aulas, el 8.7 %. No fue significativa y negativa para ninguna.

La pregunta que se hacen los financiadores es ¿Cuál es la probabilidad de que el valor del índice diferencia sea mayor a cero? El gráfico de la Figura 18 muestra la probabilidad por aula. Es una mirada distinta a lo mismo que expresa la Figura 17. Si bien todas las probabilidades son mayores a cero, un financiador esperaría que la mayoría tuvieran una probabilidad cercana al 100 %.

Por tanto, queda respondida la segunda pregunta: La intervención no logra intrínsecamente efectos uniformes en todas las instituciones educativas. Depende en gran medida de las condiciones propias de cada contexto.

Se procede a contestar la tercera pregunta, acerca de la brecha por sexo, por medio del *primer modelo jerárquico de dos niveles*.

Los gráficos de oruga de la Figura 19 presentan por cada aula la media de los coeficientes $\beta_0$ y $\beta_1$ y sus correspondientes intervalos de confianza a un nivel del 95 %.

Hay tanto valores positivos y significativos como negativos y significativos. Por tanto, una primera conclusión es que la intervención no logra resultados uniformes en todos los establecimientos educativos. Se puede interpretar en el sentido de que el resultado depende de las condiciones internas de cada institución educativa, e incluso de cada aula, y no es efectivo por sí mismo.

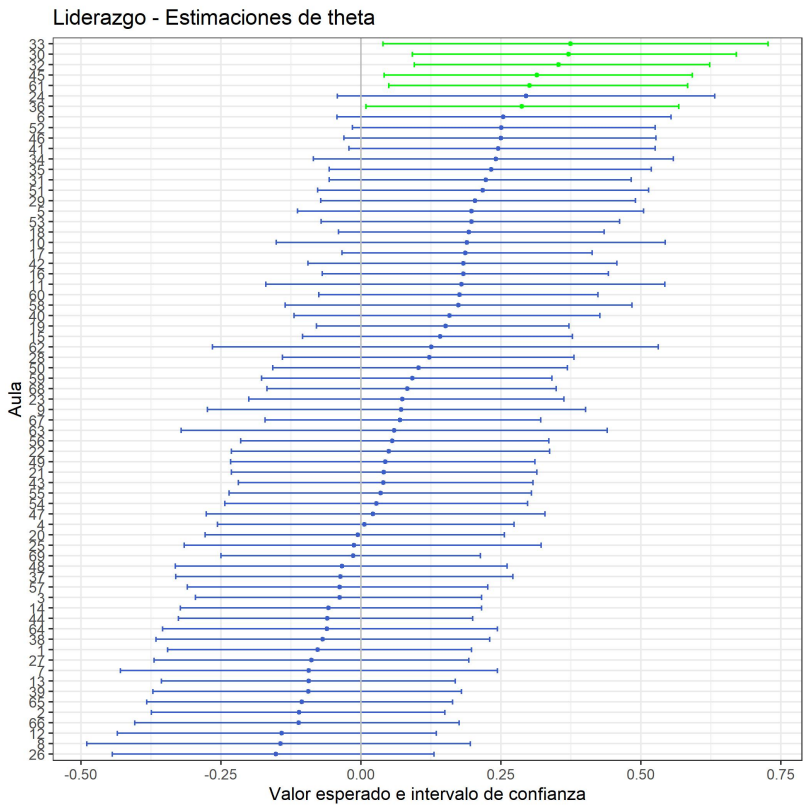



Figura 17: Representación de la media de la diferencia, por aula, para el índice Liderazgo

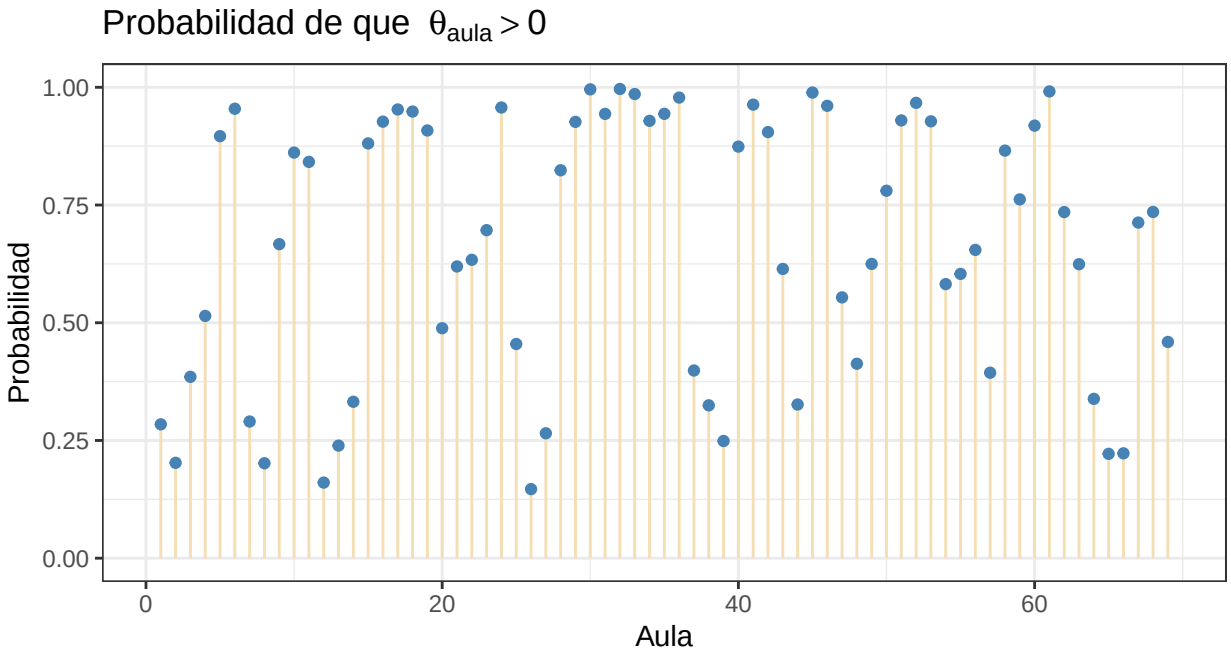



Figura 18: Probabilidad de que la media sea mayor a cero para el índice Liderazgo

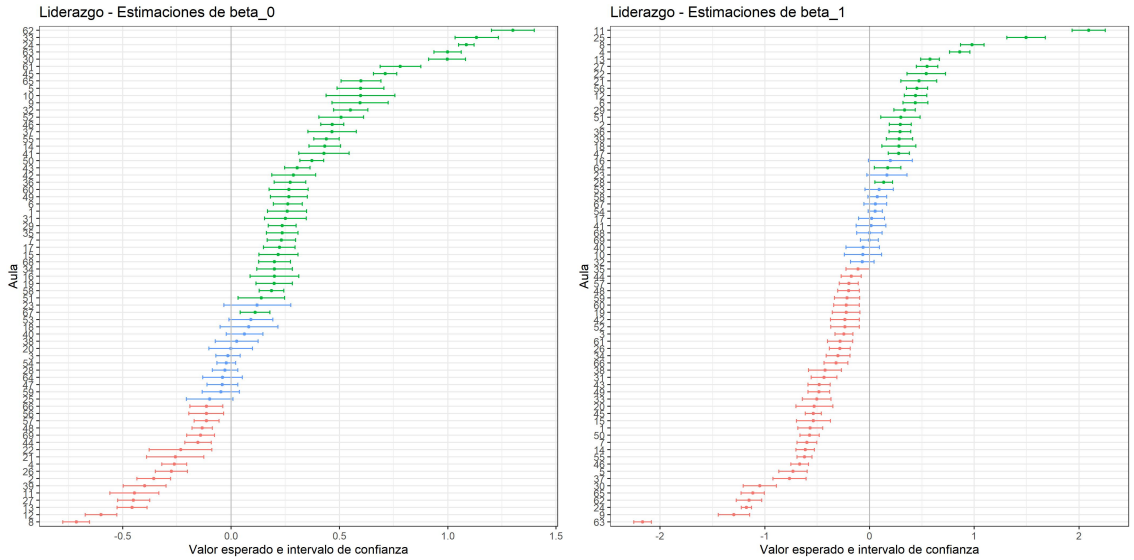



Figura 19: Coeficientes del modelo jerárquico de dos niveles para sexo

El parámetro $\beta_0$ (panel izquierdo de la Figura 19) es el coeficiente asociado a la variable indicadora que representa al sexo de referencia: *Masculino*. Permite identificar para cuántas aulas el efecto fue positivo para los hombres, es decir, en cuantos la medición indica que aumentó el índice de Liderazgo efectivo: 39, el 56.5 %. Y para cuantas disminuyó: 17, el 24.6 %.

$\beta_1$ (panel derecho de la Figura 19) es el coeficiente asociado a la variable indicadora que presenta la diferencia de las mujeres respecto a los hombres. La mayoría son coeficientes significativos, por ende, hay una brecha entre sexos.

La situación de las mujeres debe ser evaluada sumando $\beta_0$ y $\beta_1$

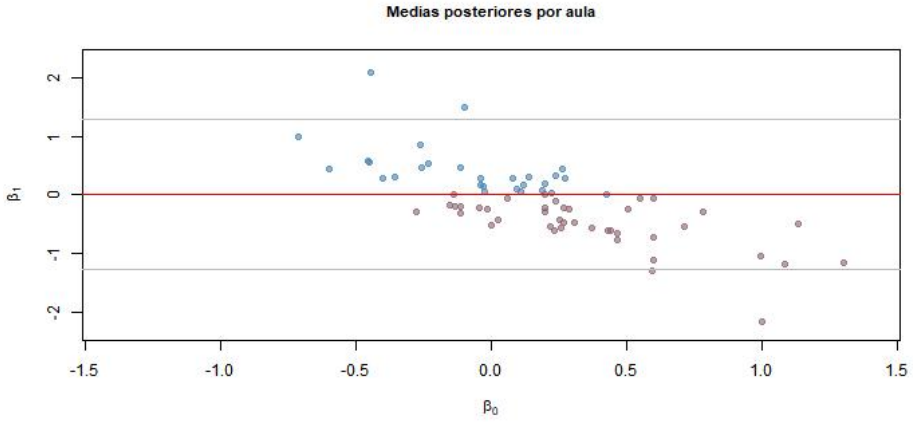



Figura 20: Representación de la totalidad de los coeficientes para el índice Liderazgo

La Figura 20, que presenta la comparación de los coeficientes $\beta_0$ y $\beta_1$, muestra en rojo una línea horizontal a la altura del cero. Los puntos sobre la línea roja indican ausencia de brecha entre sexos. Los puntos por debajo implican ventaja del sexo masculino frente al femenino. Por encima,

lo contrario. Las lineas horizontales grises indican ± una desviación estándar respecto al momento inicial.

Hay brecha tanto a favor como en contra de las mujeres. Esto indica que si bien el fútbol como deporte ha sido tradicionalmente masculino, la intervención logra realizar una práctica orientada a valores que independiza el posible sesgo que se pudiera temer en un inicio. Pero la brecha es significativa en la mayoría de los casos, así que la intervención en valores no logra evitar que haya brecha hacia alguno de los dos sexos.

La Figura 21 presenta la distribución posterior de los betas para un aula:

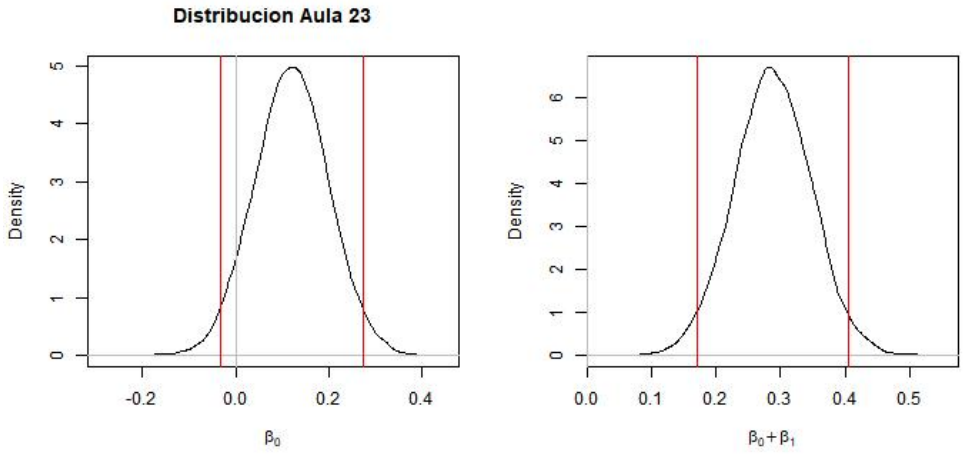
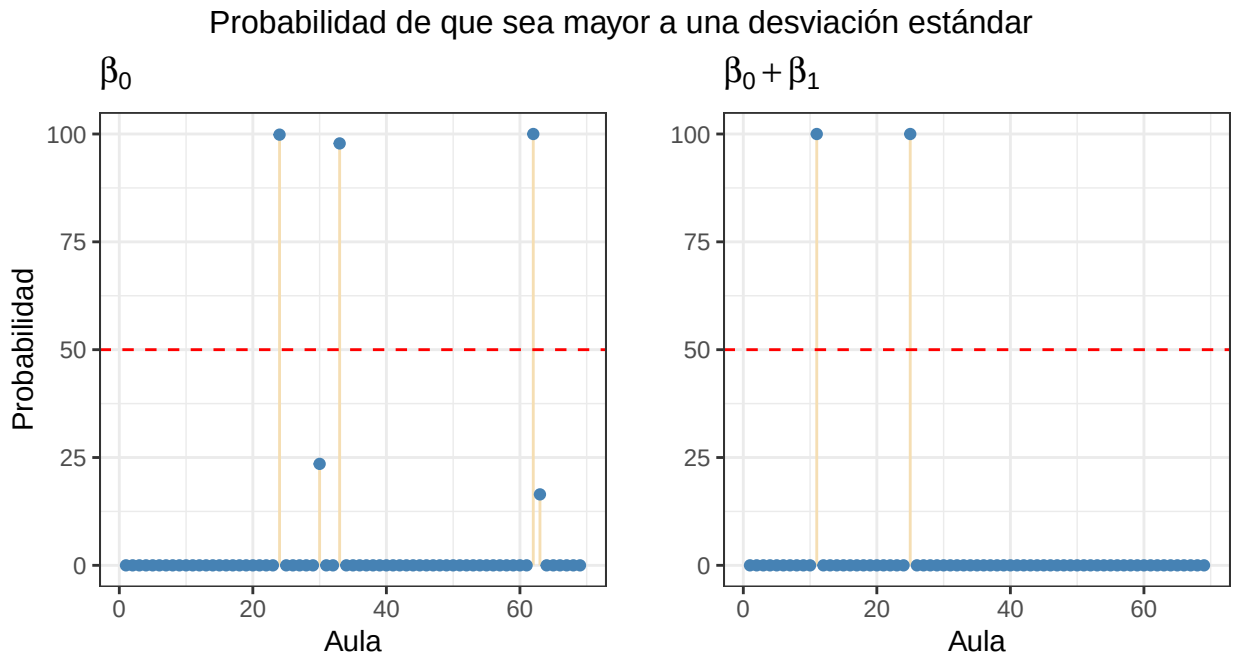



Figura 21: Distribución posterior de los betas para un aula

Para el aula presentada en el gráfico la probabilidad de que sea mayor a cero, para $\beta_0$, es 93.43 % y de que $\beta_0 + \beta_1$ sea mayor a cero es 100 %.

No se conoce cuánto es lo normal que avance, o retroceda, un estudiante adolescente sin ninguna intervención en cuanto a competencias socioemocionales o sociogrupales. Por esa razón, no se puede juzgar a partir de los datos si la intervención ha tenido efectos positivos. Tan sólo es posible exponer lo que se registra con el objeto de avanzar en el conocimiento de este tipo de intervenciones. Además, el cambio no mide el efecto atribuible a la intervención, sino el causado tanto por factores endógenos como exógenos, entre ellos la intervención. El Instituto colombiano para la Evaluación de la Educación (ICFES) utiliza la desviación estándar como un baremo factible para comparar diferentes evaluaciones[4].

[4]Por ejemplo, el *Análisis de las diferencias de género en el desempeño de estudiantes colombianos en matemáticas y lenguaje* de la serie Estudios del ICFES lo utiliza, así como el informe *Tras la excelencia docente: cómo mejorar la calidad de la educación para todos los colombianos* realizado por prominentes investigadores para la Fundación Compartir




Se observa que la probabilidad de que sea mayor a una desviación estándar sólo está presente para tres aulas. También podría ocurrir haya habido un retroceso. La probabilidad de que hayan retrocedido al menos una desviación estándar es cero.

La pregunta acerca de la diferencia por grado se responde por medio del *segundo modelo jerárquico de dos niveles*.

En el segundo modelo jerárquico de dos niveles $\beta_0$ es el coeficiente asociado a la variable indicadora del grado de referencia: *Sexto*. Por esa razón, sólo es posible calcularlo para los establecimientos educativos que tienen grado sexto (1). Por ende, $\beta_1$ es el coeficiente asociado a la variable indicadora que representa la diferencia del grado *séptimo* respecto al sexto, en consecuencia sólo tiene sentido calcularlo para los establecimientos que tienen ambos grados, sexto y séptimo (2 en total).

Aquellos establecimientos educativos que tienen grado séptimo y no sexto presentan una varianza demasiado amplia.

En las figuras 22 y 23 se observan promedios positivos y negativos significativos indicando que la intervención no afecta de manera uniforme a todos los establecimientos educativos, aún para un mismo grado. En grado sexto se observa una proporción del 0 % de establecimientos educativos con efectos positivos. En grado séptimo dicha proporción se debe calcular como resultado de sumar $\beta_0$ y $\beta_1$. Es del 0 %.

Se aplica el mismo razonamiento para los coeficientes $\beta_2$ y $\beta_3$, con 3 y 3 establecimientos educativos respectivamente. La Figura 23 presenta los gráficos correspondientes.

La limitación de comparar sólo establecimientos educativos en donde se tengan ambos grados implica una restricción en el diseño de este tipo de análisis. Una opción es obligar que en el muestreo de establecimientos educativos siempre haya un aula con el grado que se vaya a tomar como referencia. Otra opción es realizar el análisis agrupando grados. Por ejemplo, agrupar grados sexto y séptimo para que sirvan de referencia y grados octavo y noveno como segunda categoría. De todos modos implica asegurarse de tener un aula en cada una de las categorías en cada institución educativa que

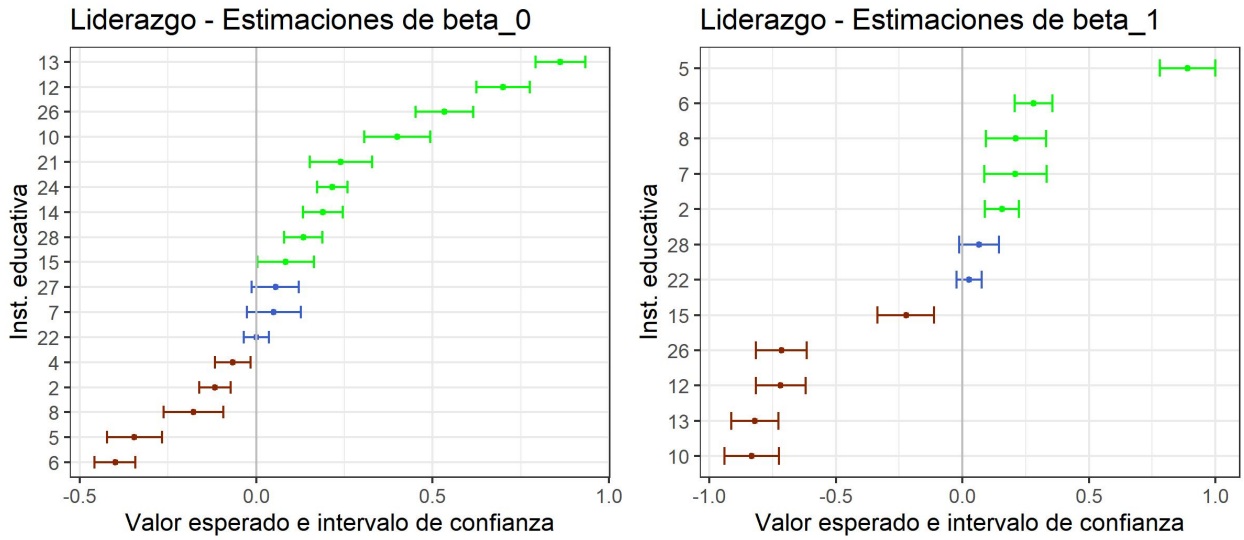



Figura 22: Dos primeros coeficientes del segundo modelo jerárquico de dos niveles

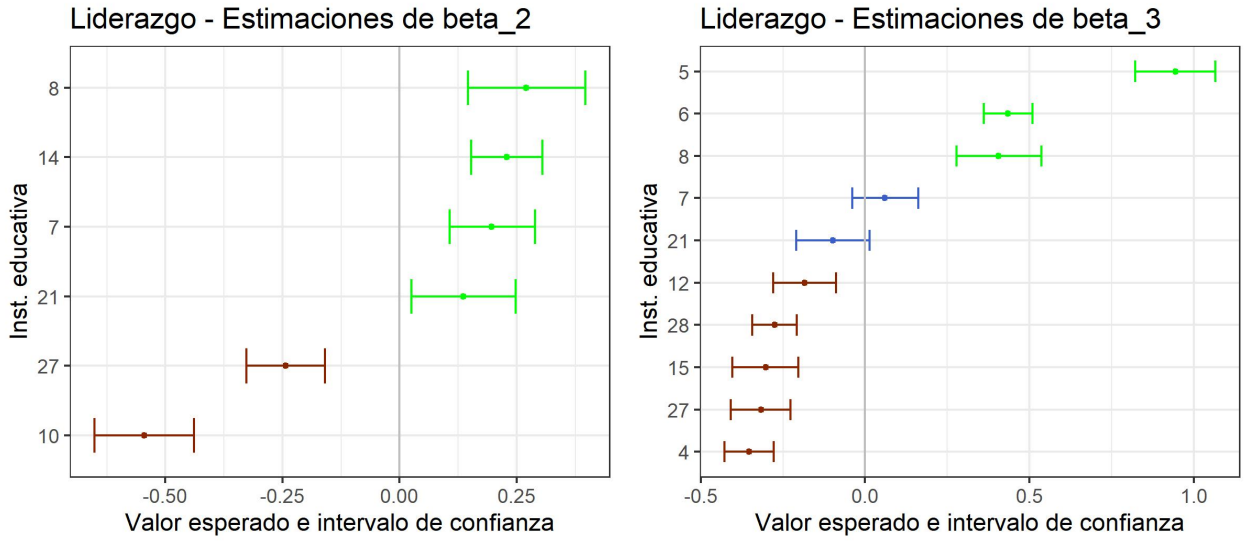



Figura 23: Restantes coeficientes del segundo modelo jerárquico de dos niveles

se incluya en el análisis.

En grado octavo se observa una proporción del 0 % de establecimientos educativos con efectos positivos, en grado noveno dicha proporción es del 0 %.

Los datos no muestran que sea un poco más efectivo aplicar la intervención en unos grados que en otros.

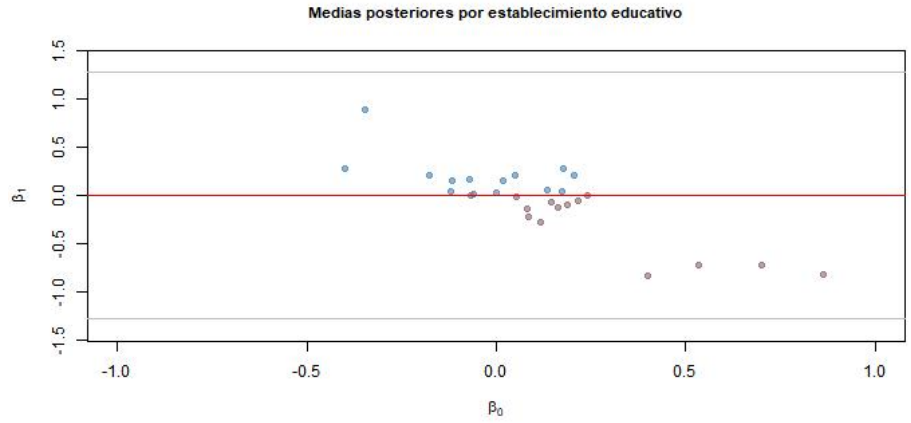



Figura 24: Coeficiente por establecimiento educativo para grado séptimo

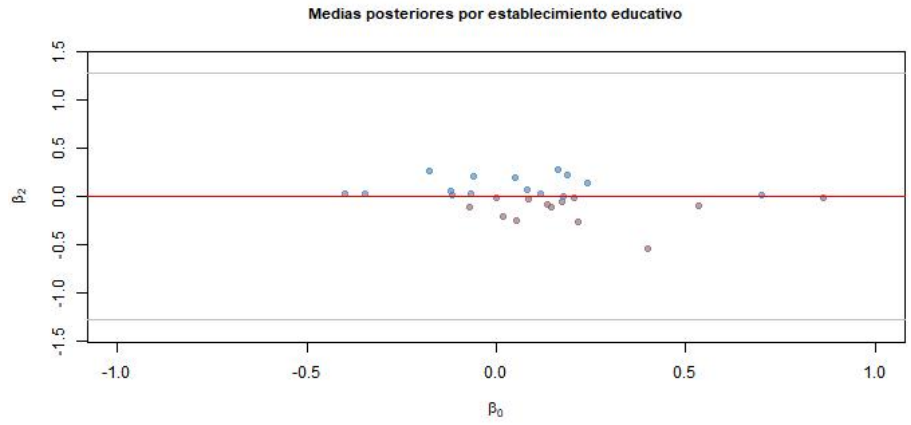



Figura 25: Coeficiente por establecimiento educativo para grado octavo

Las figuras 24, 25 y 26 presentan la comparación de los coeficientes de grado séptimo, octavo y noveno, respectivamente, respecto a grado sexto. En los tres se observan coeficientes positivos y negativos respecto al coeficiente de grado sexto, pero ninguno tiene los valores por fuera de una desviación estándar respecto al momento inicial. Implica que la intervención no es necesariamente mejor para unos grados que para otros.

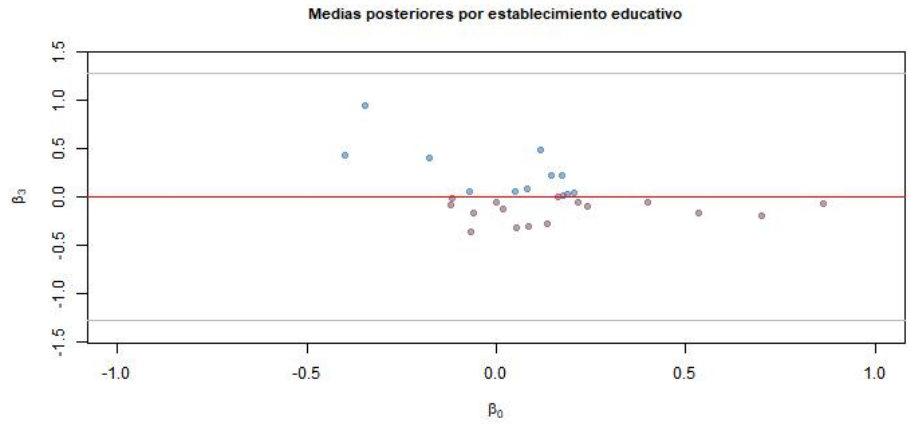



Figura 26: Coeficiente por establecimiento educativo para grado noveno

## 6.2 Habilidades comunicativas

Como ya se mencionó, la primera y segunda pregunta se responden con el *modelo jerarquíco de tres niveles*.

El gráfico de oruga de la Figura 27 identifica el valor medio y el intervalo de confianza del índice diferencia. La intervención fue significativa y positiva para nueve aulas, el 13 %. No disminuyó para ninguna.

La pregunta que se hacen los financiadores es ¿Cuál es la probabilidad de que el valor del índice diferencia analizado sea mayor a cero? El gráfico de la Figura 28 muestra la probabilidad por aula. Es una mirada distinta a lo mismo que expresa la Figura 27. Si bien todas las probabilidades son mayores a cero, un financiador esperaría que la mayoría tuvieran una probabilidad cercana al 100 %.

Por tanto, queda respondida la segunda pregunta: La intervención no logra intrínsecamente efectos uniformes en todas las instituciones educativas. Depende en gran medida de las condiciones propias de cada contexto.

Se procede a contestar la tercera pregunta, acerca de la brecha por sexo, por medio del *primer modelo jerárquico de dos niveles*.

Los gráficos de oruga de la Figura 29 presentan por cada aula la media de los coeficientes $\beta_0$ y $\beta_1$ y sus correspondientes intervalos de confianza a un nivel del 95 %.

Hay tanto valores positivos y significativos como negativos y significativos. La intervención no logra resultados uniformes en todos los establecimientos educativos.

El parámetro $\beta_0$ es el coeficiente asociado a la variable indicadora que representa al sexo de referencia: *Masculino*. Permite identificar para cuántas aulas el efecto fue positivo para los hombres, es decir, en cuantos la medición indica que aumentó el índice de Habilidad efectivo: 30, el 43.5 %. Y para cuantas disminuyó: 26, el 37.7 %. Llaman la atención tres aulas que está tienen un valor muy distinto de las restantes.

$\beta_1$ es el coeficiente asociado a la variable indicadora que presenta la diferencia de las mujeres respecto a los hombres. La mayoría son coeficientes significativos, por ende, hay una brecha entre sexos.

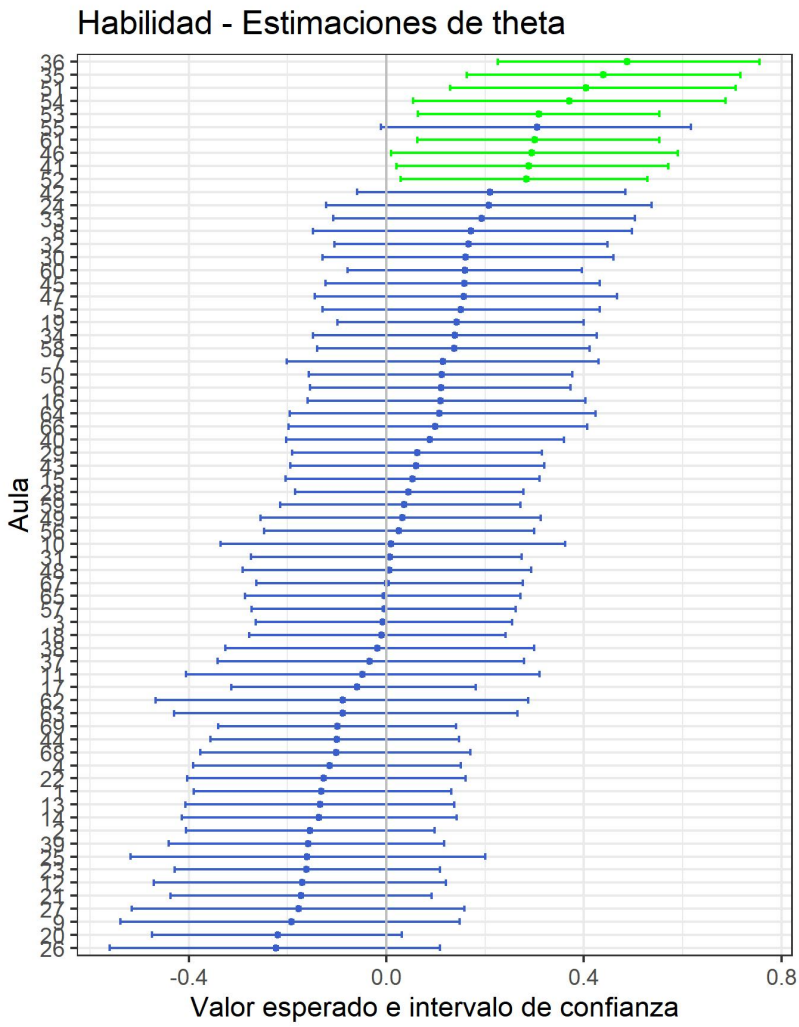



Figura 27: Representación de la media de la diferencia, por aula, para el índice Habilidad

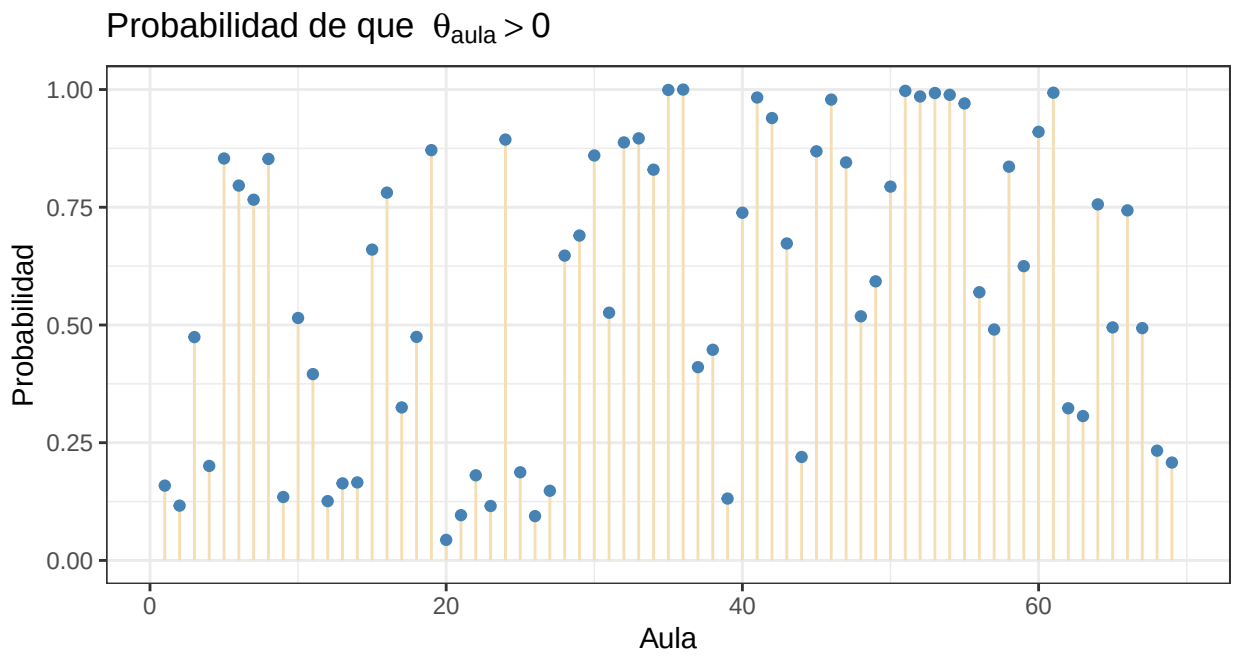



Figura 28: Probabilidad de que la media sea mayor a cero para el índice Habilidad

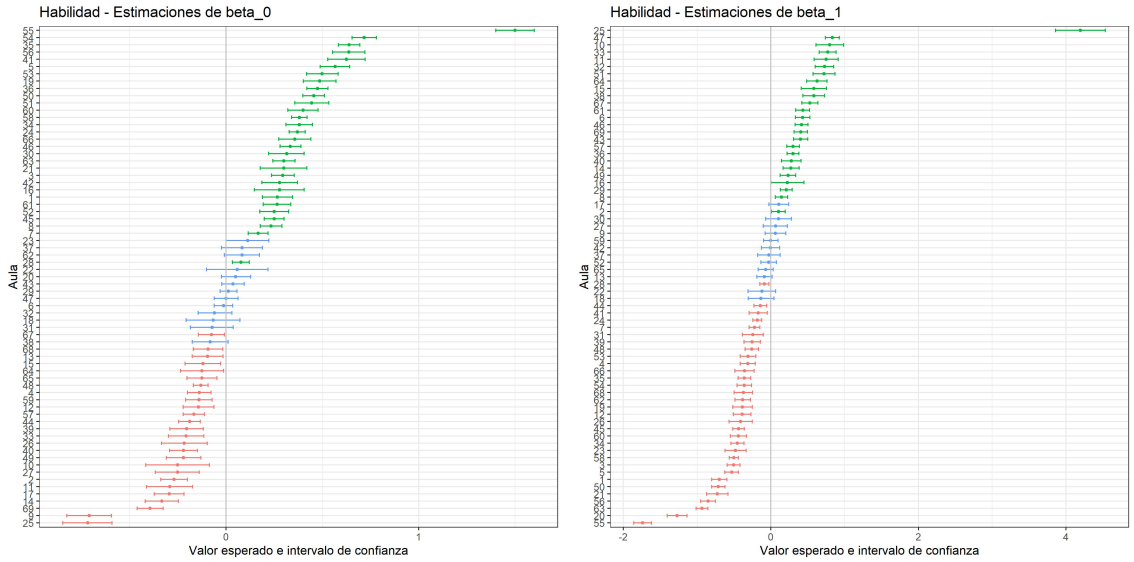



Figura 29: Coeficientes del modelo jerárquico de dos niveles para sexo

La situación de las mujeres debe ser evaluada sumando $\beta_0$ y $\beta_1$

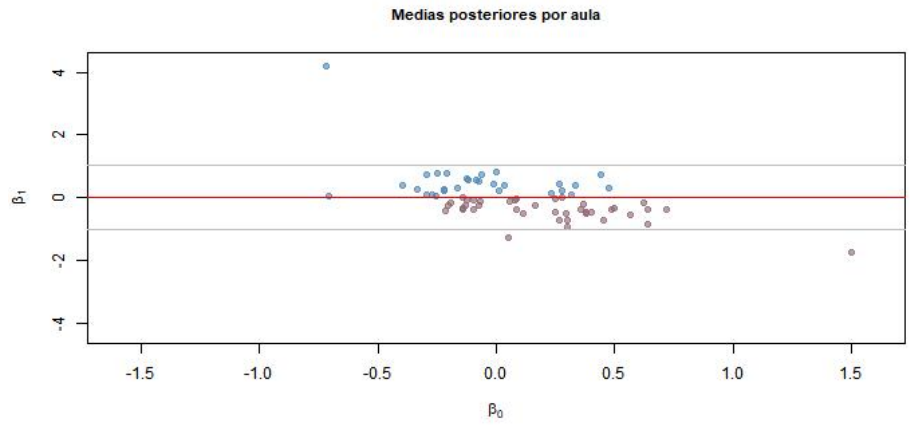



Figura 30: Representación de la totalidad de los coeficientes para el índice Habilidad

La Figura 30, que presenta la comparación de los coeficientes $\beta_0$ y $\beta_1$, muestra en rojo una línea horizontal a la altura del cero. Los puntos sobre la línea roja indican ausencia de brecha entre sexos. Los puntos por debajo implican ventaja del sexo masculino frente al femenino. Por encima, lo contrario. Las lineas horizontales grises undican $\pm$ una desviación estándar respecto al momento inicial.

Hay brecha tanto a favor como en contra de las mujeres, así que la intervención en valores no logra evitar que haya brecha hacia alguno de los dos sexos. Hay tres aulas con valores mayores a una desviación estándar. Dos le dan ventaja al grupo de varones y una al grupo femenino.

La Figura 31 presenta la distribución posterior de los betas para un aula:

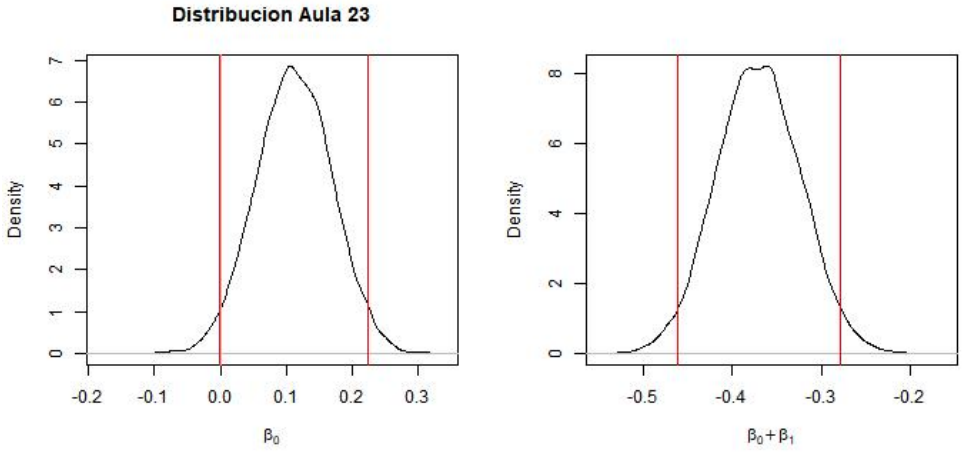
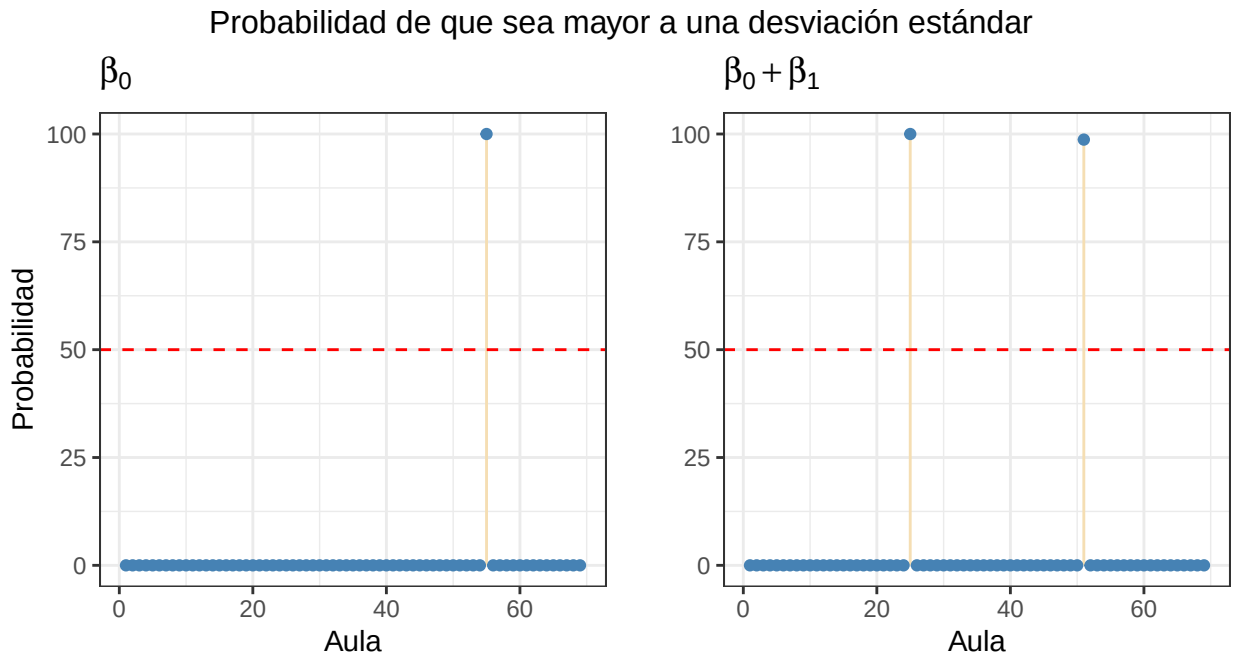



Figura 31: Distribución posterior de los betas para un aula

Para el aula presentada en el gráfico la probabilidad de que sea mayor a cero, para $\beta_0$, es 97.46 % y de que $\beta_0 + \beta_1$ sea mayor a cero es 0 %. Es una aula que presenta una gran brecha entre sexos.




Se observa que la probabilidad de que sea mayor a una desviación estándar sólo está presente para dos aulas. También podría ocurrir haya habido un retroceso. La probabilidad de que hayan retrocedido al menos una desviación estándar es cero, excepto para un aula.

La pregunta acerca de la diferencia por grado se responde por medio del *segundo modelo jerárquico de dos niveles*.

En el segundo modelo jerárquico de dos niveles $\beta_0$ es el coeficiente asociado a la variable indicadora del grado de referencia: *Sexto*. Por esa razón, sólo es posible calcularlo para los establecimientos educativos que tienen grado sexto (0). Por ende, $\beta_1$ es el coeficiente asociado a la variable indicadora que representa la diferencia del grado *séptimo* respecto al sexto, en consecuencia, sólo tiene sentido calcularlo para los establecimientos que tienen ambos grados, sexto y séptimo (0 en total).

En las figuras 32 y 33 se observan promedios positivos y negativos significativos indicando que la intervención no afecta de manera uniforme a todos los establecimientos educativos, aún para un mismo grado. En grado sexto se observa una proporción del NaN % de establecimientos educativos con efectos positivos. En grado séptimo dicha proporción se debe calcular como resultado de sumar $\beta_0$ y $\beta_1$. Es del NaN %.

Se aplica el mismo razonamiento para los coeficientes $\beta_2$ y $\beta_3$, con 0 y 0 establecimientos educativos respectivamente. La Figura 33 presenta los gráficos correspondientes.

La limitación de comparar sólo establecimientos educativos en donde se tengan ambos grados implica una restricción en el diseño de este tipo de análisis. Una opción es obligar que en el muestreo de establecimientos educativos siempre haya un aula con el grado que se vaya a tomar como referencia. Otra opción es realizar el análisis agrupando grados. Por ejemplo, agrupar grados sexto y séptimo para que sirvan de referencia y grados octavo y noveno como segunda categoría. De todos modos implica asegurarse de tener un aula en cada una de las categorías en cada institución educativa que se incluya en el análisis.

En grado octavo se observa una proporción del NaN % de establecimientos educativos con efectos

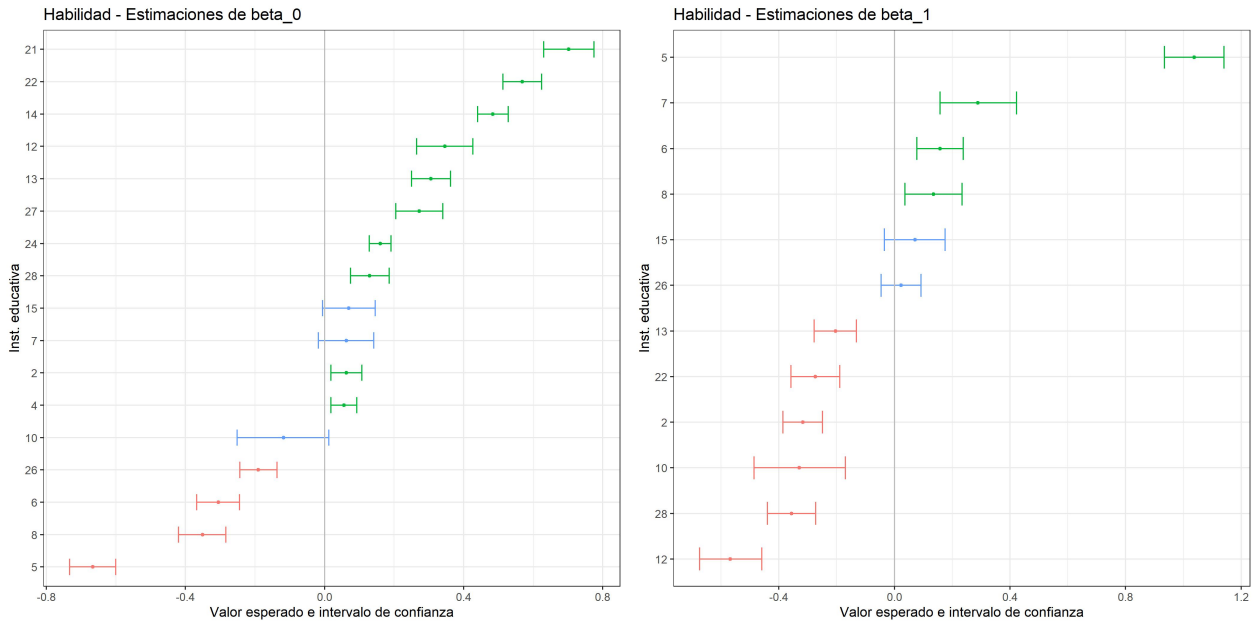



Figura 32: Dos primeros coeficientes del segundo modelo jerárquico de dos niveles

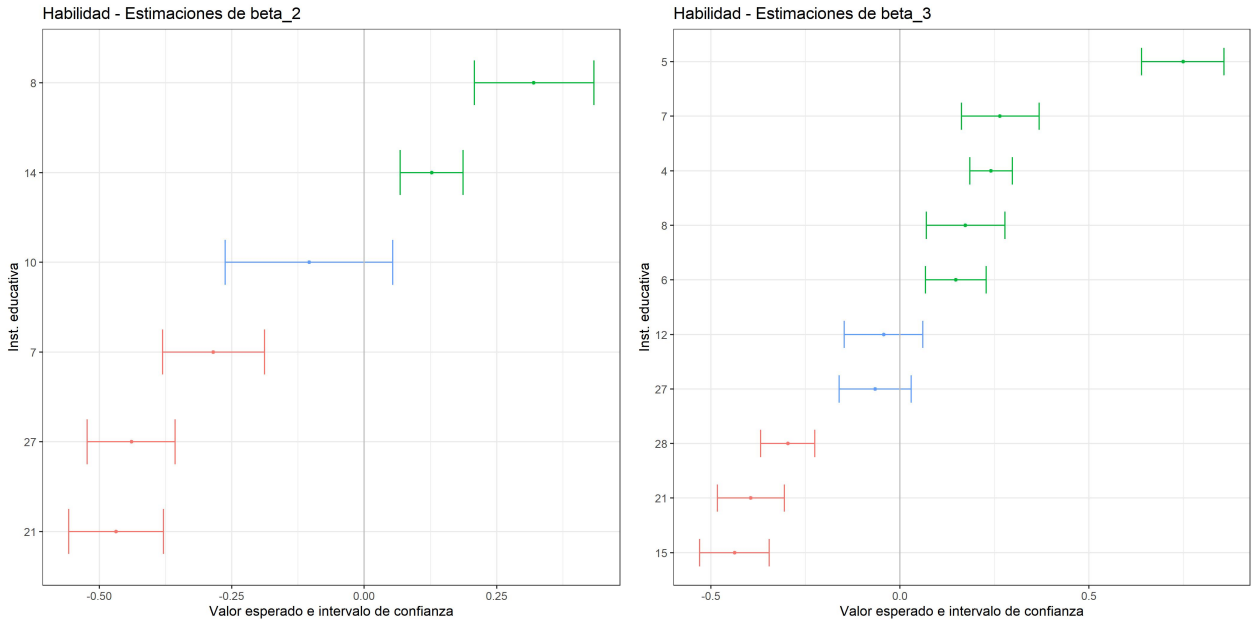



Figura 33: Restantes coeficientes del segundo modelo jerárquico de dos niveles

positivos, en grado noveno dicha proporción es del NaN %.

Los datos no muestran un patrón en el que se identifique que sea un poco más efectivo aplicar la intervención en unos grados que en otros. De hecho, en los dos grados extremos se obtiene un mayor porcentaje que en los intermedios.

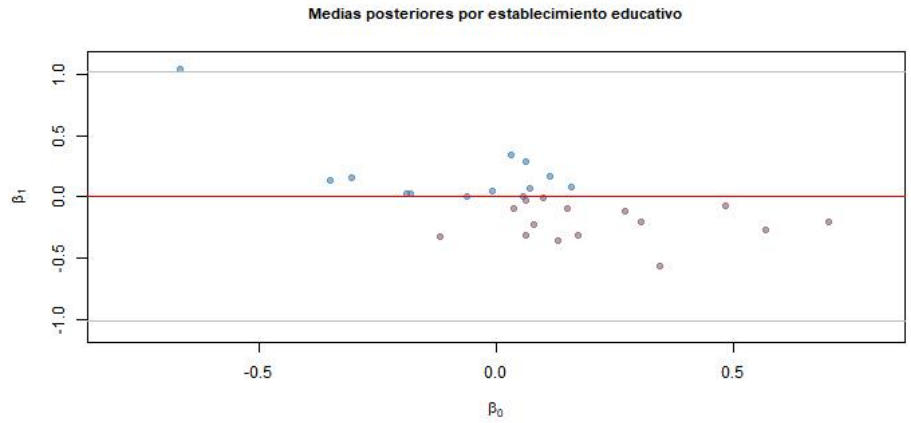



Figura 34: Coeficiente por establecimiento educativo para grado séptimo

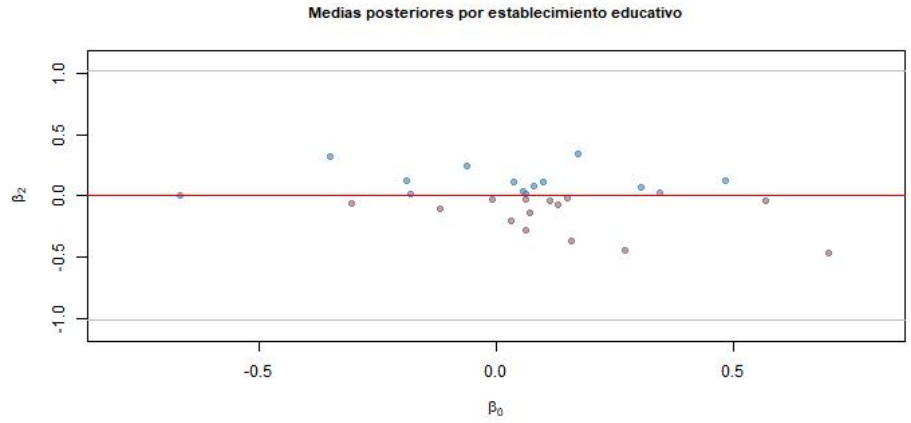



Figura 35: Coeficiente por establecimiento educativo para grado octavo

Las figuras 34, 35 y 36 presentan la comparación de los coeficientes de grado séptimo, octavo y noveno, respectivamente, respecto a grado sexto. En los tres se observan coeficientes positivos y negativos respecto al coeficiente de grado sexto, pero, excepto uno de grado séptimo, ninguno tiene los valores por fuera de una desviación estándar respecto al momento inicial. La intervención no es necesariamente mejor para unos grados que para otros.

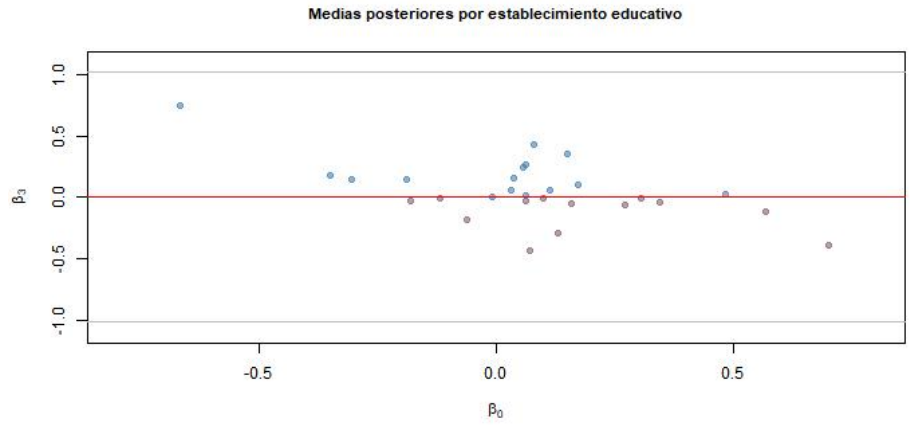



Figura 36: Coeficiente por establecimiento educativo para grado noveno

### 6.3 Respeto por sí mismo y por los demás

Como ya se mencionó, la primera y segunda pregunta se responden con el *modelo jerarquíco de tres niveles*.

El gráfico de oruga de la Figura 37 identifica el valor medio y el intervalo de confianza del índice diferencia analizado. La intervención fue significativa y negativa para ocho aulas, el 11.6 %. No aumentó para ninguna.

La pregunta que se hacen los financiadores es ¿Cuál es la probabilidad de que el valor del índice diferencia sea mayor a cero? El gráfico de la Figura 38 muestra la probabilidad por aula. Es una mirada distinta a lo mismo que expresa la Figura 37. La mayoría de las probabilidades son menores al 50 %. Para este índice, la intervención fue desafortunada.

Por tanto, queda respondida la segunda pregunta: La intervención no logra intrínsecamente efectos uniformes en todas las instituciones educativas. Depende en gran medida de las condiciones propias de cada contexto.

Se procede a contestar la tercera pregunta, acerca de la brecha por sexo, por medio del *primer modelo jerárquico de dos niveles*.

Los gráficos de oruga de la Figura 39 presentan por cada aula la media de los coeficientes $\beta_0$ y $\beta_1$ y sus correspondientes intervalos de confianza a un nivel del 95 %.

Hay tanto valores positivos y significativos como negativos y significativos. La intervención no logra resultados uniformes en todos los establecimientos educativos.

El parámetro $\beta_0$ es el coeficiente asociado a la variable indicadora que representa al sexo de referencia: *Masculino*. Permite identificar para cuántas aulas el efecto fue positivo para los hombres, es decir, en cuantos la medición indica que aumentó el índice de Respeto efectivo: 22, el 31.9 %. Y para cuantas disminuyó: 38, el 55.1 %. Llaman la atención tres aulas que está tienen un valor muy distinto de las restantes.

$\beta_1$ es el coeficiente asociado a la variable indicadora que presenta la diferencia de las mujeres respecto a los hombres. La mayoría son coeficientes significativos, por ende, hay una brecha entre sexos.

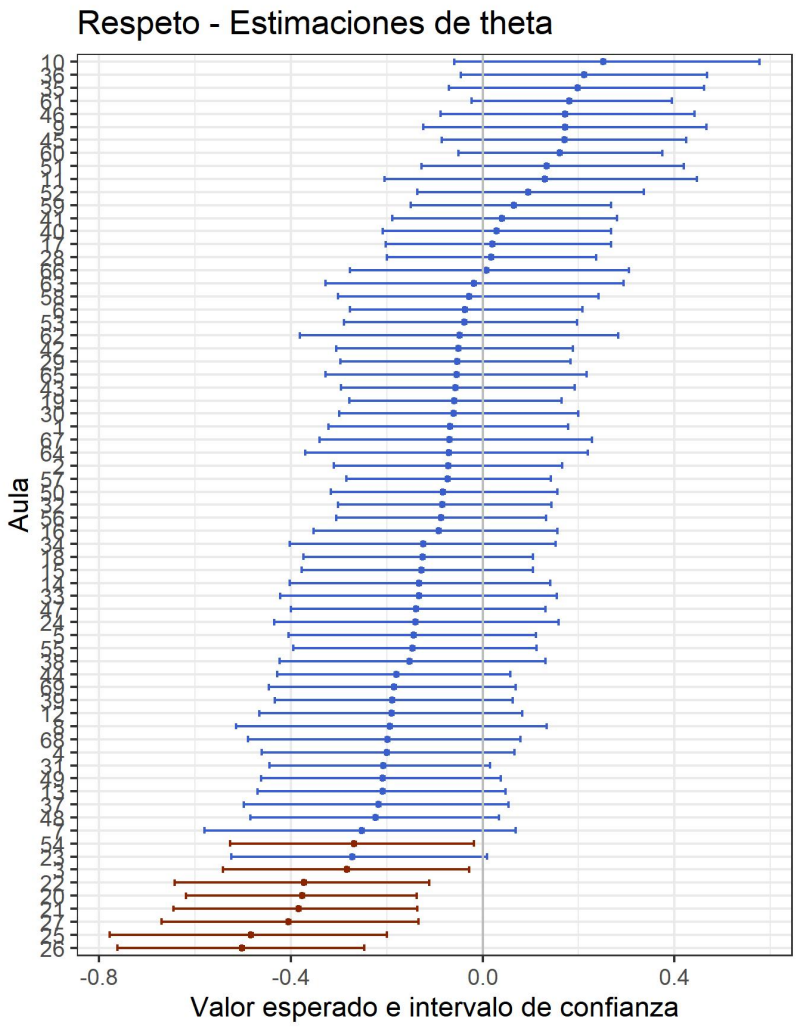



Figura 37: Representación de la media de la diferencia, por aula, para el índice Respeto

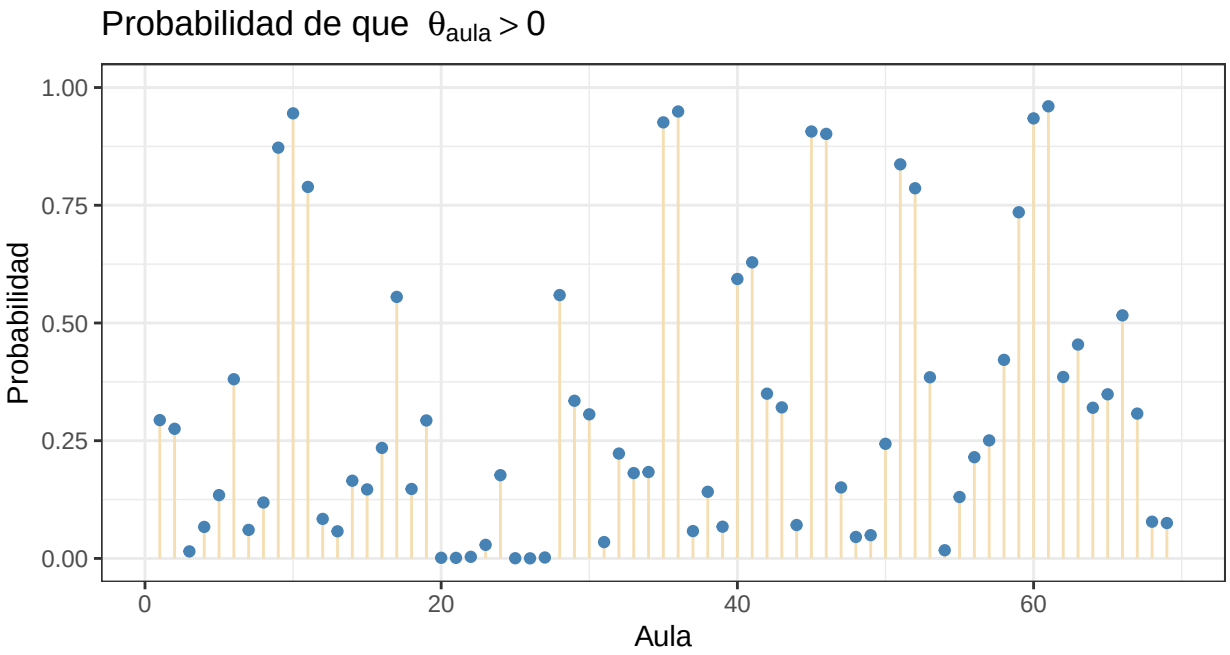



Figura 38: Probabilidad de que la media sea mayor a cero para el índice Respeto

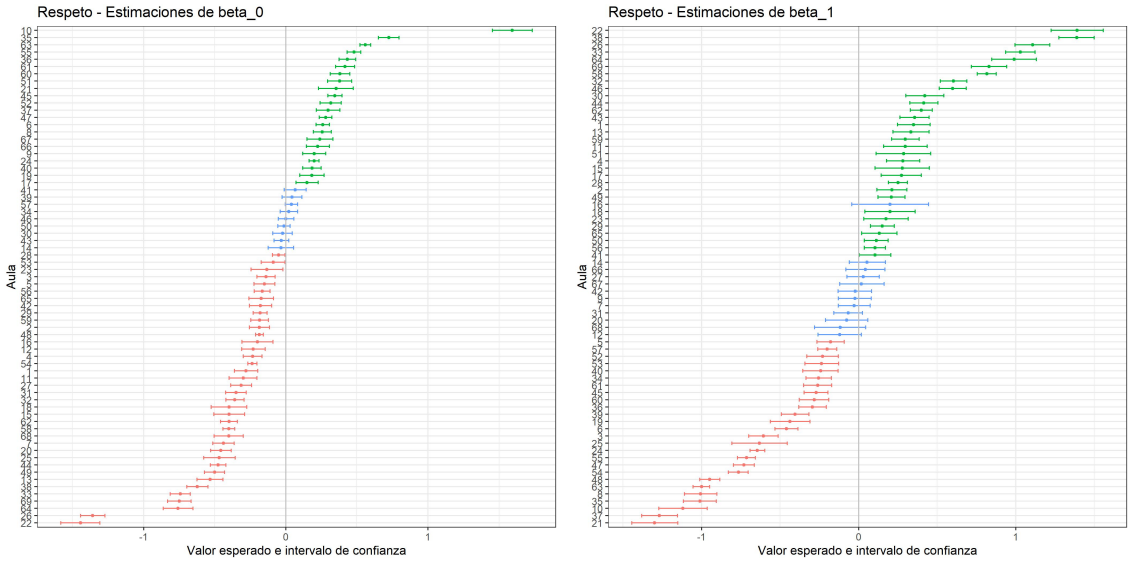



Figura 39: Coeficientes del modelo jerárquico de dos niveles para sexo

La situación de las mujeres debe ser evaluada sumando $\beta_0$ y $\beta_1$

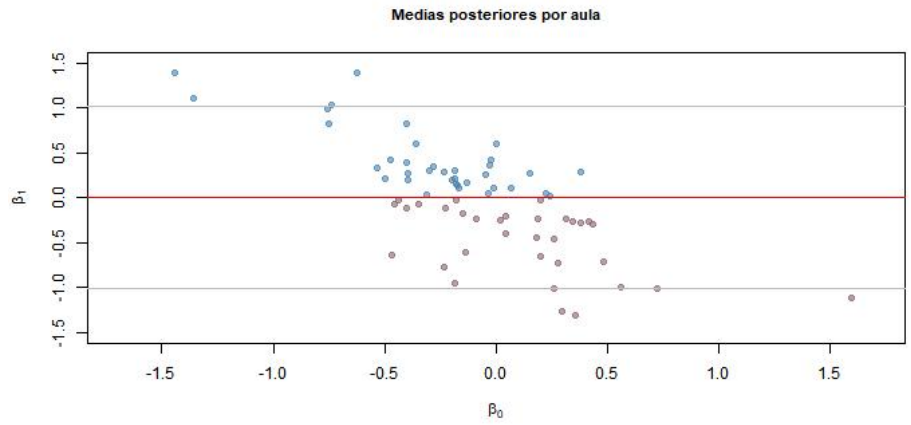



Figura 40: Representación de la totalidad de los coeficientes para el índice Respeto

La Figura 40, que presenta la comparación de los coeficientes $\beta_0$ y $\beta_1$, muestra en rojo una línea horizontal a la altura del cero. Los puntos sobre la línea roja indican ausencia de brecha entre sexos. Los puntos por debajo implican ventaja del sexo masculino frente al femenino. Por encima, lo contrario. Las lineas horizontales grises undican $\pm$ una desviación estándar respecto al momento inicial.

Hay brecha tanto a favor como en contra de las mujeres, así que la intervención en valores no logra evitar que haya brecha hacia alguno de los dos sexos. Hay una decena de aulas con valores mayores a una desviación estándar. Unas le dan ventaja al grupo de varones y otras al grupo femenino.

La Figura 41 presenta la distribución posterior de los betas para un aula:

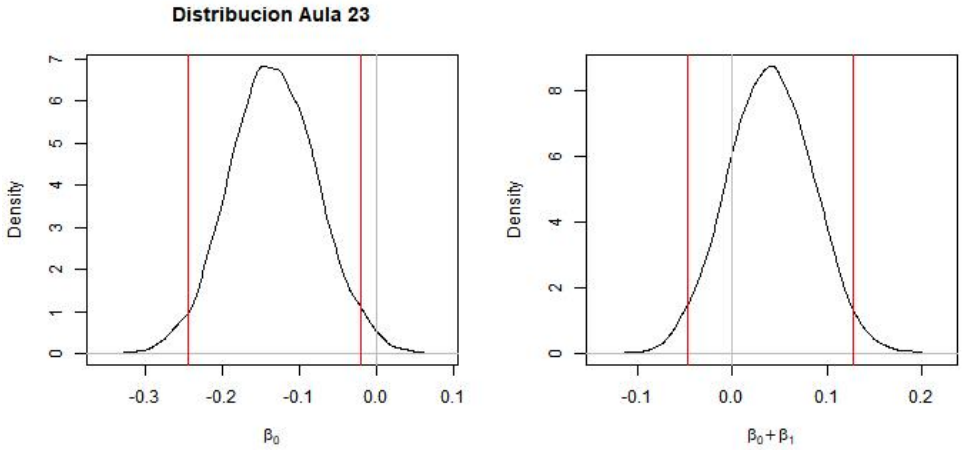
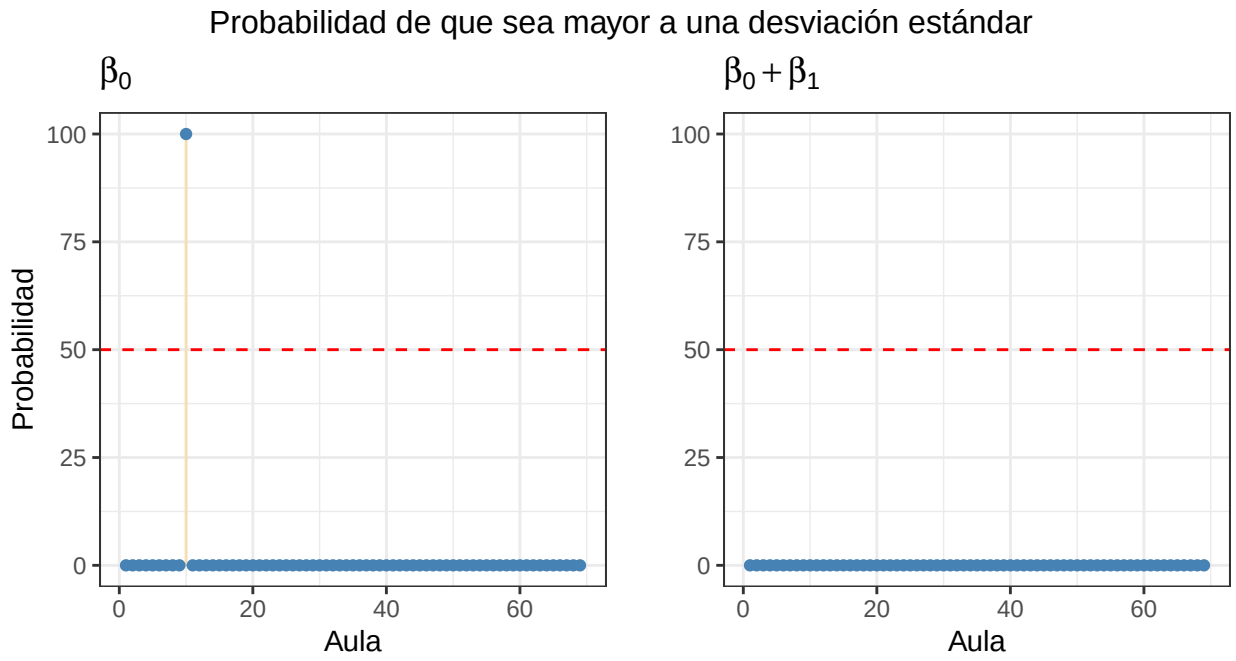
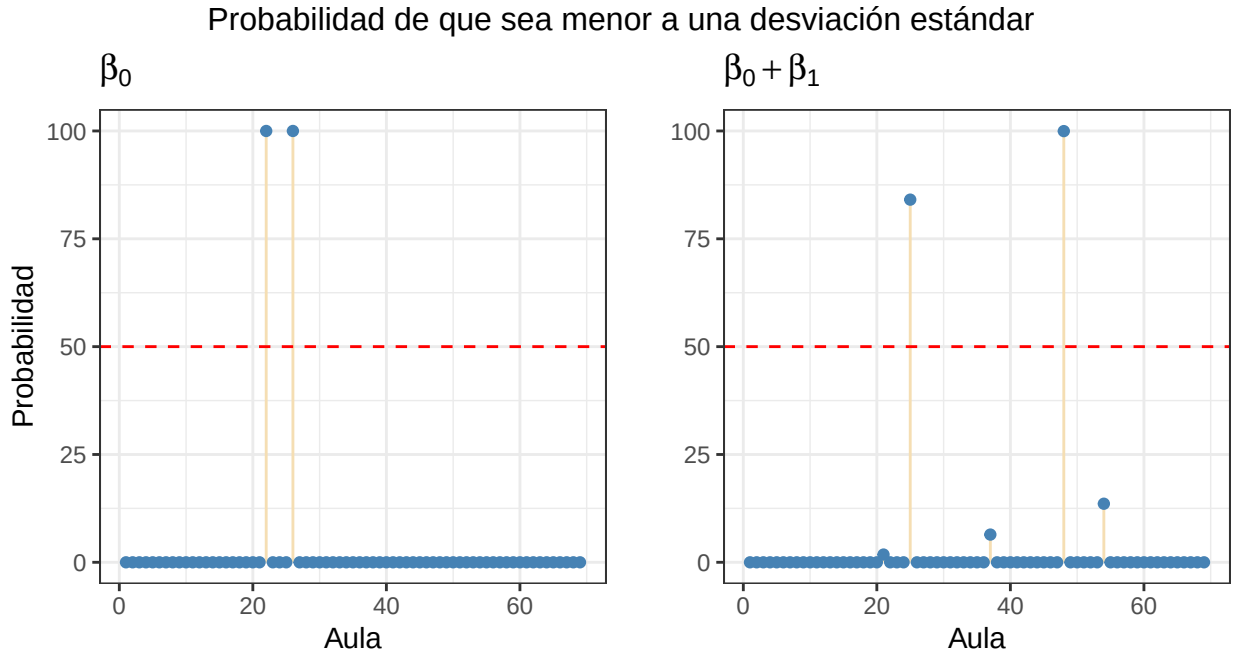



Figura 41: Distribución posterior de los betas para un aula

Para el aula presentada en el gráfico la probabilidad de que sea mayor a cero, para $\beta_0$, es 1.01 % y de que $\beta_0 + \beta_1$ sea mayor a cero es 81.22 %. Es una aula que presenta una gran brecha entre sexos, en contra del sexo masculino.




Se observa que la probabilidad de que sea mayor a una desviación estándar sólo está presente para una única aula. También podría ocurrir haya habido un retroceso. La probabilidad de que hayan retrocedido al menos una desviación estándar es alrededor del 30 % para una decena de aulas para los hombres, y de alrededor del 40 % para ocho aulas respecto a las mujeres.




La pregunta acerca de la diferencia por grado se responde por medio del *segundo modelo jerárquico de dos niveles*.

En el segundo modelo jerárquico de dos niveles $\beta_0$ es el coeficiente asociado a la variable indicadora del grado de referencia: *Sexto*. Por esa razón, sólo es posible calcularlo para los establecimientos

educativos que tienen grado sexto (0). Por ende, $\beta_1$ es el coeficiente asociado a la variable indicadora que representa la diferencia del grado *séptimo* respecto al sexto, en consecuencia, sólo tiene sentido calcularlo para los establecimientos que tienen ambos grados, sexto y séptimo (0 en total).

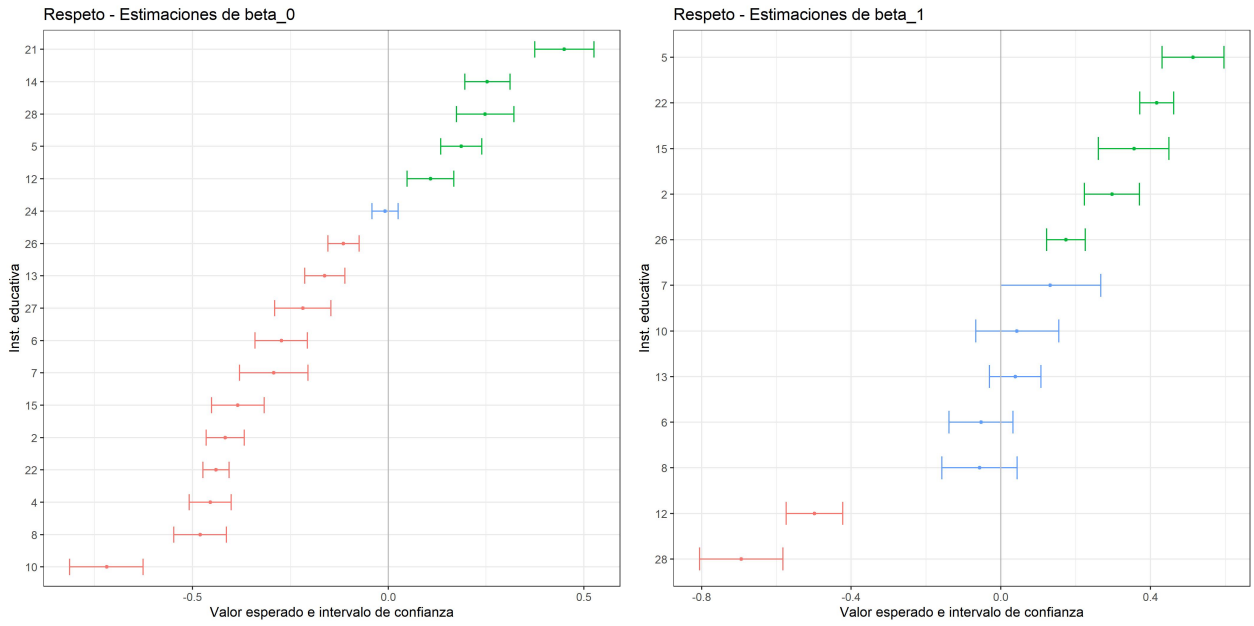



Figura 42: Dos primeros coeficientes del segundo modelo jerárquico de dos niveles

En las figuras 42 y 43 se observan promedios positivos y negativos significativos indicando que la intervención no afecta de manera uniforme a todos los establecimientos educativos, aún para un mismo grado. En grado sexto se observa una proporción del NaN % de establecimientos educativos con efectos positivos. En grado séptimo dicha proporción se debe calcular como resultado de sumar $\beta_0$ y $\beta_1$. Es del NaN %.

Se aplica el mismo razonamiento para los coeficientes $\beta_2$ y $\beta_3$, con 0 y 0 establecimientos educativos respectivamente. La Figura 43 presenta los gráficos correspondientes.

La limitación de comparar sólo establecimientos educativos en donde se tengan ambos grados implica una restricción en el diseño de este tipo de análisis. Una opción es obligar que en el muestreo de establecimientos educativos siempre haya un aula con el grado que se vaya a tomar como referencia. Otra opción es realizar el análisis agrupando grados. Por ejemplo, agrupar grados sexto y séptimo para que sirvan de referencia y grados octavo y noveno como segunda categoría. De todos modos implica asegurarse de tener un aula en cada una de las categorías en cada institución educativa que se incluya en el análisis.

En grado octavo se observa una proporción del NaN % de establecimientos educativos con efectos positivos, en grado noveno dicha proporción es del NaN %.

Los datos no muestran un patrón en el que se identifique que sea un poco más efectivo aplicar la intervención en unos grados que en otros. Grado octavo presenta un porcentaje similar al de grado sexto, y grado séptimo similar en magnitud al grado noveno.

Las figuras 44, 45 y 46 presentan la comparación de los coeficientes de grado séptimo, octavo y

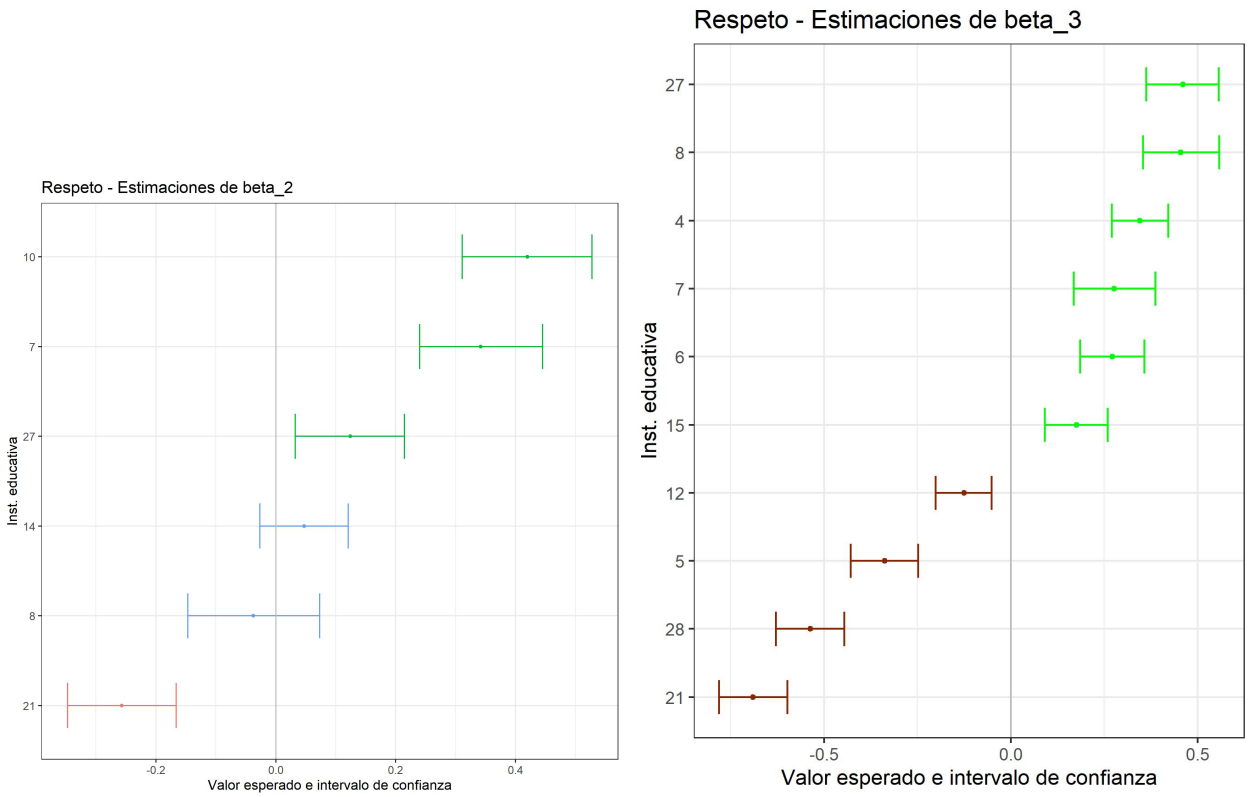



Figura 43: Restantes coeficientes del segundo modelo jerárquico de dos niveles

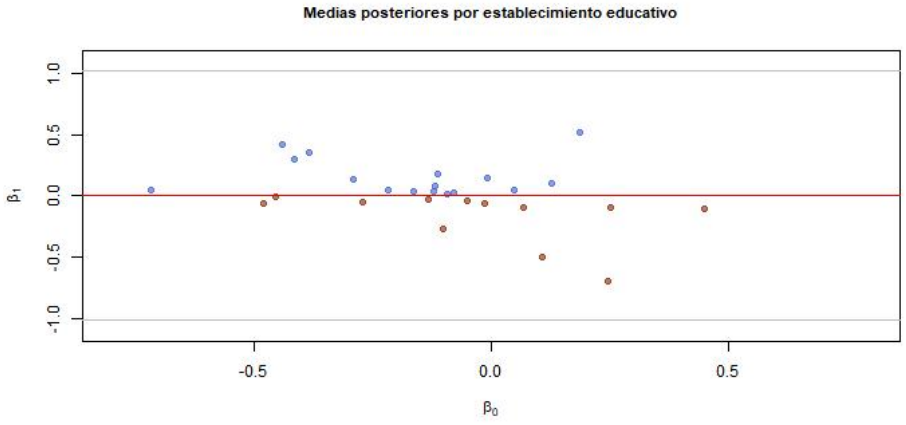



Figura 44: Coeficiente por establecimiento educativo para grado séptimo

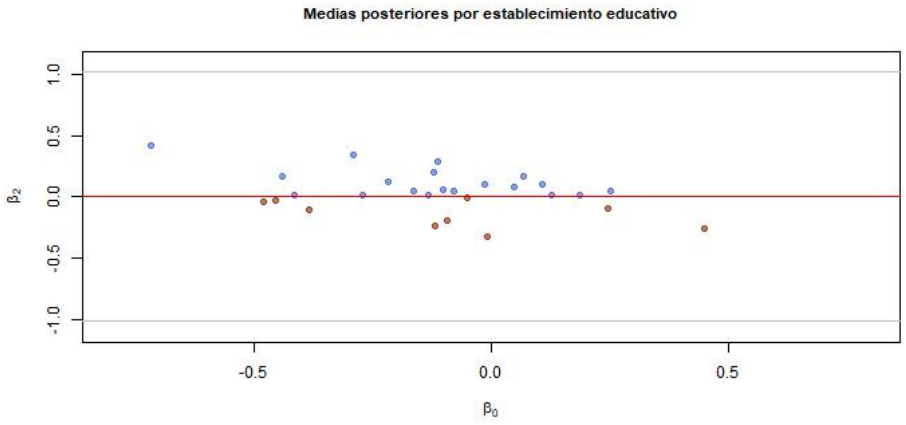



Figura 45: Coeficiente por establecimiento educativo para grado octavo

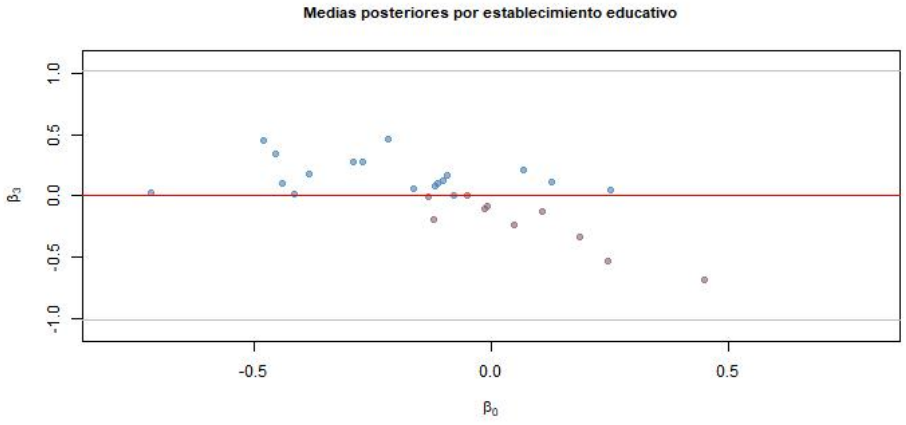



Figura 46: Coeficiente por establecimiento educativo para grado noveno

noveno, correspondientemente, respecto a grado sexto. En los tres se observan coeficientes positivos y negativos respecto al coeficiente de grado sexto, pero, ninguno tiene valores por fuera de una desviación estándar respecto al momento inicial. La intervención no es necesariamente mejor para unos grados que para otros.

# 7. Conclusiones

Hay dos tipos de conclusiones: las asociadas con la metodología estadística y las asociadas con la aplicación.

## 7.1. La metodología estadística

Los modelos jerárquicos planteados fueron efectivos en permitir comparar la brecha entre sexos o entre grados. El modelo jerárquico de tres niveles mostró un comportamiento de la mezcla muy bueno, permitiendo hacer uso de los resultados de la diferencia promedio en cada aula de cada institución educativa, y de la varianza correspondiente a cada institución educativa. Pero presentó una mezcla pobre y un número efectivo de muestras bajo para los parámetros que corresponden a la varianza de la media. Hoff [11] escribe en el numeral 11.5 de su texto que los MCMC de los modelos jerárquicos pueden sufrir de este problema. Congdon [3] menciona que tanto Givens and Hoeting (2012) como Browne (2004) proponen realizar una transformación orientada a centrar las variables para reducir la correlación posterior de las distribuciones posteriores e incrementar el tamaño efectivo de las muestras (pag 318). Es una estrategia que no fue viable aplicar como mejora por cuanto los índices ya están bastante centrados alrededor de cero y su desviación estándar cercana a uno.

La aplicación se realizó sobre el sector educativo, pero hay otros sectores que son naturalmente jerárquicos y en los que se puede aplicar esta misma técnica: salud pública, ecosistemas o estructuras empresariales.

## 7.2. La aplicación

Los resultados de la intervención dejan claro que la metodología depende en gran medida del contexto donde se aplique. Para los tres indices se obtuvieron brechas entre sexos, tanto en favor de uno como del otro, indicando que per se no hay discriminación hacia un sexo u otro. La brecha la genera el contexto, el cual contiene muchos cofactores: el docente, la dinámica del grupo, el porcentaje de estudiantes de un sexo u otro, la cultura institucional, la cultura social del ambiente del que provienen, ...

La diferencia entre grados también es en ambas direcciones, sin dejar entrever claramente que haya un grado en que sea más conveniente intervenir para ninguno de los índices analizados.

Aunque la intervención muestre algunas evidencias de mejora en el desempeño en los índices medidos, esta eficiencia debe evaluarse si es significativa en términos de la inversión de recursos y tiempo aplicados a la intervención. Quizá se deben establecer estrategias diferenciales de acuerdo a las características de la institución, los grados y las condiciones soeciodemográficas de los estudiantes, entre otras.

Los resultados cercanos a cero se pueden explicar por el corto periodo de intervención. Casey y Goodyear mencionan que la literatura que rodea el desarrollo del aprendizaje de los estudiantes en dominios físicos, cognitivos, sociales y afectivos se centra principalmente en responder a la pregunta de si funciona, en lugar de preguntar cuáles son los beneficios para los estudiantes y su aprendizaje a lo largo del tiempo. Mencionan que Kirk (2010), basándose en el trabajo de Ennis (1999), sostuvo que unidades de instrucción que duran entre cuatro y seis lecciones no permiten que el aprendizaje progrese más allá del nivel más básico. Que un creciente cuerpo de investigación sugiere que se necesitan varias unidades iniciales antes que los estudiantes aprendan a aprender de

esta manera. Así que el énfasis en estudios cortos es una limitación de las investigaciones[2]. Ellas se refieren al Aprendizaje Cooperativo, pero bien se puede extender a otro tipo de pedagogías. Añaden que el Aprendizaje Cooperativo, como otros modelos pedagógicos, a menudo se ha aplicado solo dentro de un plan de estudios más amplio, dentro de actividades múltiples. Hace referencia a que la intervención es una de los varios proyectos que se planean para un mismo periodo académico. Ambas limitaciones son pertinentes para el caso de la intervención analizada: la intervención se realizó en 10 sesiones, dejando las demás sesiones para desarrollar otras actividades planeadas por parte del Área de Educación Física, Recreación y Deporte, y dicho número de sesiones se queda corto para consolidar unos resultados bajo una metodología que es extraña al discurrir pedagógico usual y que por ende requiere de un tiempo de aprendizaje acerca de cómo desarrollarlo, tanto para los alumnos como para los docentes.

Casey y Goodyear también exponen que hay limitaciones en el enfoque de cada estudio de los revisados por ellas. Más particularmente, mencionan la brevedad de muchas intervenciones (menos de 6 semanas en algunos casos) y la falta de detalle acerca de si se mantuvo fidelidad al modelo[2].

Velázquez escribe que "... Smith y Goc Karp (1997) ya demostraron que programas puntuales de juegos cooperativos pueden no resultar tan eficaces si no se prolongan en el tiempo y van acompañados de otras acciones en las clases. Es más, en grupos poco habituados a cooperar pueden manifestarse algunos comportamientos negativos, como la tendencia de algunos estudiantes a actuar individualmente, incluso perjudicando las respuestas cooperativas de otros compañeros, o la comparación de resultados entre personas o grupos (Lavega, Planas y Ruíz, 2014)" pag. 278[19].

Si bien para los tres índices se presentaron resultados menores a una desviación estándar, hay indicios de que el índice de *Respeto por sí mismo y hacia los demás* tuvo un comportamiento tendiente a negativo. Sin intención de querer explicarlo, denota que la técnica logra revelar cambios pequeños.

En la exposición se presentaron gráficas que expresaban la probabilidad por aula de que la media estuviera por encima o por debajo de una desviación estándar. Es el tipo de flexibilidad que proporciona el paradigma Bayesiano. Una vez se obtienen las cadenas, se pueden realizar multitud de preguntas sin tener que definir y calcular modelos nuevos.

La probabilidad expresada en dichas gráficas coincide en la mayoría de los casos con el concepto de probabilidad que manejan los tomadores de decisiones de este tipo de proyectos. Una ventaja del paradigma Bayesiano que ya se mencionó en la Sección 2.

Los análisis por aula o por establecimiento educativo son cercanos a las necesidades de los operadores quienes recolectan mucha información cualitativa que pueden contrastar con la cuantitativa para determinar hipótesis acerca de las razones por las que funcionó la intervención en unos contextos y no en otros.

### 7.3 Futuros estudios

Congdon[3] recuerda que otro objetivo importante de los modelos jerárquicos es estudiar la partición de la varianza. Por ejemplo, ¿qué proporción de las variación del desempeño en valores sociogrupales se debe a las características de los establecimientos educativos, y cuánto se debe a las características de su contexto familiar y social.

El modelamiento presentado es susceptible de mejorarse. Algunas opciones son:

a. Usar previas sobre los coeficientes $\beta$ dado que la cantidad de cofactores puede ser grande.

b. Se puede usar alguna otra previa para los parámetros de varianza, con el objeto de lograr que se puedan hacer inferencias respecto a la variabilidad.

c. El MCMC puede tomar tiempo en correr y probar otras alternativas puede ser costoso en tiempo, así que también se puede usar cálculo variacional para explorar la distribución posterior (Variational Inference A Review for Statisticians), o la técnica de Approximate Bayesian Computation (ABC), usando Integrated Nested Laplace Approximation (INLA). Y para seleccionar los modelos, además de los criterios de información y la validación cruzada se pueden usar Factores de Bayes (sección 1.8 de Jackman[13]).

Además de intervalos de credibilidad, se pueden hacer pruebas de hipótesis (sección 1.8 de Jackman[13]).

Dado que la variable respuesta no tiene que ser necesariamente normal, desde un punto de vista no paramétrico se puede modelar la distribución que asume la variable respuesta por medio de una previa a través de un proceso de Dirichlet (Capítulo 23 del libro de Gelman[8]).

## Referencias


[1] Chaux, E. et al., *Competencias ciudadanas. De los estándares al aula. Una propuesta integral para todas las áreas académicas*, (2004) compiladores, Enrique Chaux, Juanita Lleras, Ana María Velásquez. – Bogotá: Ministerio de Educación, Universidad de los Andes, Facultad de Ciencias Sociales, Departamento de Psicología y Centro de Estudios Socioculturales e Internacionales, Ediciones Uniandes.

[2] Casey, A. y Goodyear, V., *Can Cooperative Learning Achieve the Four Learning Outcomes of Physical Education? A Review of Literature*, Journal Quest. Vol 67, Issue I (2015).

[3] Congdon, P.D., *Bayesian Hierarquical Models with applications using R*, Second Edition. (2020) CRC Press. Taylor and Francis Group. Boca raton, FL, USA

[4] Delprato, M., *Determinantes del rendimiento educativo del nivel primario aplicando la Técnica de Análisis Multinivel*, (1999) IERAL, Documentos de trabajo N° 27, Córdoba.

[5] Gastwirth, J.L, *Statistical Science in the Courtroom*, (2000) Springer. NY

[6] Gelman, A. y Hill, J., *Data analysis using Regression and Multilevel/Hierarchical Models*, (2007) Cambridge, NY.

[7] Gelman, A., *Understanding posterior p-values*, Electronic Journal of Statistics. Vol. 7 (2013) 2595–2602

[8] Gelman, A., et.al., *Bayesian data analysis*, Third Edition, (2014) Taylor and Francis Group

[9] Gutierrez, A, *Estrategias de muestreo, diseño de encuestas y estimación de parámetros.* (2015) Ediciones de la Universidad Santo Tomás.

[10] Gutierrez, F.G., *Conceptos y clasificación de las capacidades físicas.* Revista de investigación Cuerpo, cultura y movimiento, 1(1), 77-86 (2009). Ediciones USTA.

[11] Hoff, P.D., *A First Course in Bayesian Statistical Methods*, (2009) Springer, USA.

[12] Holt, N., *Positive Youth Development through Sport*, (2016) 2nd edition, Edited by Nicholas L. Holt. Routledge.

[13] Jackman, S., *Bayesian Analysis for the Social Sciences*, (2009) John Wiley & Sons.

[14] McCoach, D.B., *Hierarquical Linear Models*, en The Reviwer's guide to quantitative methods in the social sciences. (2010) Edited by Hancock and Mueller. Routledge. New York.

[15] Mosteller, F. y Boruch, R., *Evidence matters*, (2002) Brookings Institution Press.

[16] Murillo Tordecilla, F.J, *Los modelos multinivel como herramienta para la investigación educativa*, (2012) Magis, Revista Internacional de Investigación en Educación, 1(1).

[17] Organización de Estados Americanos, *Deporte y educación ciudadana: ¿Cómo educar en valores y prácticas democráticas a través del deporte?*, (2013) Boletín sobre Educación y Democracia. Departamento de Desarrollo Humano, Educación y Empleo SEDI. Programa Interamericano sobre educación en valores y prácticas democráticas. Organización de Estados Americanos OEA. Sep.

[18] Snijders, T.A.B y Bosker, R.J., *Multilevel Analysys, An Introduction to Basic and Advanced Multilevel Modeling*, (2012) 2nd edition, Sage.

[19] Velázquez Callado, C., *El enfoque de pedagogía como pieza clave en la transformación social* En Educación Física y Pedagogía Crítica: Propuestas para la transformación personal y social. Eloísa Lorente Catalán y Daniel Martos García, compiladores. Ediciones de la Universidad de Lleida, Valencia. España. 2018.

[20] Wasserstein, R.L. y Lazar, N.A., *The ASA's statement on p-values: context, process, and purpose* (2016) The American Statistical Association.

## Anexo A. Inferencia de las distribuciones condicionales conjuntas del segundo modelo: jerárquico de tres niveles

Se desarrolla la verosimilitud de la ecuación (6) para un aula cualquiera j de una institución educativa cualquiera k:

$$p(y_{ijk} \mid resto) = \prod_{i=1}^{n_{jk}} p(y_{ijk} \mid \theta_{jk}, \sigma_k^2) = \prod_{i=1}^{n_{jk}} \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2}\left(\frac{y_{ijk}-\theta_{jk}}{\sigma_k}\right)^2} \tag{44a}$$

$$= (2\pi\sigma_k^2)^{-(n_{jk}/2)} \exp\left[ -\frac{1}{2} \sum_{i=1}^{n_{jk}} \left( \frac{y_{ijk} - \theta_{jk}}{\sigma_k} \right)^2 \right] \tag{44b}$$

$$\propto (2\pi\sigma_k^2)^{-(n_{jk}/2)} \exp\left[ -\frac{1}{2\sigma_k^2} \sum_{i=1}^{n_{jk}} \left( y_{ijk} - \theta_{jk} \right)^2 \right] \tag{44c}$$

$$= (2\pi\sigma_k^2)^{-(n_{jk}/2)} \exp\left[ -\frac{1}{2\sigma_k^2} \sum_{i=1}^{n_{jk}} \left( y_{ijk}^2 - 2\theta_{jk} y_{ijk} + \theta_{jk}^2 \right) \right] \tag{44d}$$

$$= (2\pi\sigma_k^2)^{-(n_{jk}/2)} \exp\left[ -\frac{1}{2\sigma_k^2} \Big( \sum_{i=1}^{n_{jk}} y_{ijk}^2 - 2\theta_{jk} \sum_{i=1}^{n_{jk}} y_{ijk} + n_{jk}\theta_{jk}^2 \Big) \right] \tag{44e}$$

La distribución posterior, es decir, la actualización de las creencias acerca de los parámetros una vez se hallan observado los datos es $p(\theta_{jk}, \sigma_k^2 \mid y_{ijk})$.

Por la definición dada en la ecuación (7) y la independencia de los dos parámetros:

$$p(\theta_{jk} \mid \sigma_k^2) = p(\theta_{jk})$$

$$= \frac{1}{\sqrt{2\pi\tau_k^2}} \exp\left[ -\frac{1}{2\tau_k^2} (\theta_{jk} - \mu_k)^2 \right] \tag{45a}$$

$$\propto \exp\Big[ -\frac{1}{2\tau_k^2} (\theta_{jk} - \mu_k)^2 \Big], \tag{45b}$$

Se está en disposición de desarrollar la distribución posterior para cada $\theta_{jk}$:

$$p(\theta_{jk}, \sigma_k^2 \mid y_{ijk}) = p(\theta_{jk} \mid \sigma_k^2, y_{ijk}) p(\sigma_k^2 \mid y_{ijk}) \tag{46}$$

Son dos distribuciones condicionales.

Se resuelve la primera:

$$p(\theta_{jk} \mid resto) \propto p(y_{ijk} \mid \theta_{jk}, \sigma_k^2) p(\theta_{jk} \mid \sigma_k^2) \tag{47}$$

la cual es una distribución conjugada de la distribución a priori, es decir, normal. He aquí el desarrollo:

A partir de las ecuaciones (44a) y (45a) se expresa la (47):

$$
\begin{aligned}
p(\theta_{jk} \mid resto) &\propto \exp\Big[ -\frac{1}{2\sigma_k^2}\Big(\sum_{i=1}^{n_{jk}} y_{ijk}^2 - 2\theta_{jk}\sum_{i=1}^{n_{jk}} y_{ijk} + n_{jk}\theta_{jk}^2\Big)\Big]\exp\Big[ -\frac{1}{2\tau_k^2}(\theta_{jk}-\mu_k)^2\Big] \\
&= \exp\Big[ -\frac{1}{2\sigma_k^2}\Big(\sum_{i=1}^{n_{jk}} y_{ijk}^2 - 2\theta_{jk}\sum_{i=1}^{n_{jk}} y_{ijk} + n_{jk}\theta_{jk}^2\Big) - \frac{1}{2\tau_k^2}(\theta_{jk}-\mu_k)^2\Big] \\
&= \exp\Big[ -\frac{1}{2\sigma_k^2}\Big(\sum_{i=1}^{n_{jk}} y_{ijk}^2 - 2\theta_{jk}\sum_{i=1}^{n_{jk}} y_{ijk} + n_{jk}\theta_{jk}^2\Big) - \frac{1}{2\tau_k^2}\Big(\theta_{jk}^2 - 2\theta_{jk}\mu_k + \mu_k^2\Big)\Big], \quad \text{(48a)}
\end{aligned}
$$

Ordenando la expresión del exponente e ignorando el -1/2 por el momento:

$$
\theta_{jk}^2\left(\frac{1}{\tau_k^2} + \frac{n_{jk}}{\sigma_k^2}\right) - 2\theta_{jk}\left(\frac{\mu_k}{\tau_k^2} + \frac{\sum_{i=1}^{n_{jk}} y_{ijk}}{\sigma_k^2}\right) + \left(\frac{\mu_k^2}{\tau_k^2} + \frac{\sum_{i=1}^{n_{jk}} y_{ijk}}{\sigma_k^2}\right) \quad (49)
$$

Si cada multiplicando es rebautizado del siguiente modo:

$$
a = \left(\frac{1}{\tau_k^2} + \frac{n_{jk}}{\sigma_k^2}\right),
$$

$$
b = \left(\frac{\mu_k}{\tau_k^2} + \frac{\sum_{i=1}^{n_{jk}} y_{ijk}}{\sigma_k^2}\right),
$$

y

$$
c = \left(\frac{\mu_k^2}{\tau_k^2} + \frac{\sum_{i=1}^{n_{jk}} y_{ijk}}{\sigma_k^2}\right),
$$

entonces, la ecuación (48a) es $\exp\big[ -\frac{1}{2}(a\theta_{jk}^2 - 2b\theta_{jk} + c)\big]$.

Mediante manipulaciones algebraicas, y teniendo en cuenta que el término $c$ no es múltiplo de $\theta_{jk}$:

$$
\begin{aligned}
p(\theta_{jk} \mid resto) &\propto \exp\Big[ -\frac{1}{2}\Big[a\theta_{jk}^2 - 2b\theta_{jk}\Big]\Big] \quad \text{(50a)} \\
&= \exp\Big[ -\frac{1}{2}a\Big[\theta_{jk}^2 - 2b\theta_{jk}/a\Big] - \frac{1}{2}b^2/a + \frac{1}{2}b^2/a\Big] \\
&= \exp\Big[ -\frac{1}{2}a\Big[\theta_{jk}^2 - 2b\theta_{jk}/a + b^2/a^2\Big] + \frac{1}{2}b^2/a\Big] \\
&\propto \exp\Big[ -\frac{1}{2}a\Big[(\theta_{jk} - b/a)^2\Big]\Big] \\
&= \exp\Big[ -\frac{1}{2}\frac{(\theta_{jk} - b/a)^2}{1/a}\Big] \quad \text{(50b)}
\end{aligned}
$$

La ecuación (50a) se corresponde con la de una distribución normal con $E(x) = b/a$ y $Var(x) = 1/a$, lo cual concuerda con el hecho de que la previa y posterior sean conjugadas.

Sean entonces las siguientes definiciones:

$$B_{jk}^2 = \frac{1}{a} = \frac{1}{\frac{1}{\tau_k^2} + \frac{n_{jk}}{\sigma_k^2}} = \Big[\frac{\tau_k^2\sigma_k^2}{\sigma_k^2 + n_{jk}\tau_k^2}\Big] \tag{51a}$$

$$A_{jk} = \frac{b}{a} = \frac{\frac{1}{\tau_k^2}\mu_k + \frac{n_{jk}}{\sigma_k^2}\bar{y}_{jk}}{\frac{1}{\tau_k^2} + \frac{n_{jk}}{\sigma_k^2}} = \Big[\frac{1}{\tau_k^2}\mu_k + \frac{n_{jk}}{\sigma_k^2}\bar{y}_{jk}\Big]\Big[\frac{\tau_k^2\sigma_k^2}{\sigma_k^2 + n_{jk}\tau_k^2}\Big], \tag{51b}$$

que corresponden a la varianza y la media, respectivamente, de la distribución posterior del $\theta_{jk}$.

La transformación del término b desde $\sum_{i=1}^{n_{jk}} y_{ijk}$ a $n_{jk}\bar{y}_{jk}$ es importante en el sentido de recordarnos que la media se puede interpretar como una ponderación entre la esperanza de la previa y la información a posteriori. Si $A_{jk}$ y $B_{jk}^2$ dependen de $n_{jk}$, significa que la distribución posterior depende del tamaño de cada grupo, es decir, aún suponiendo que los parámetros de la distribución previa son los mismos para cada aula de una misma institución educativa, no lo son los de la distribución posterior!

Se dice entonces que la posterior de $\theta_{jk} \sim N(A_{jk}, B_{jk}^2)$ y la ecuación (50a) se reescribe entonces así:

$$p(\theta_{jk} \mid \sigma_k^2, y_{ijk}, \mu_k, \tau_k^2) \propto \exp\Bigg[ -\frac{1}{2}\frac{(\theta_{jk} - A_{jk})^2}{B_{jk}^2}\Bigg] \tag{52}$$

---

Condicionado a que $\sigma_k^2$ es único por institución educativa:

$$p(\sigma_k^2 \mid resto) \propto \prod_{j=1}^{n_k}\Bigg(\prod_{i=1}^{n_{jk}}\Big(p(y_{ijk} \mid \theta_{jk}, \sigma_k^2)\Big)\Bigg)p(\sigma_k^2) \tag{53}$$

Por la definición de la ecuación (8)

$$\begin{aligned} p(\sigma_k^2) &= \frac{\eta^\alpha}{\Gamma\Big(\alpha\Big)}(\sigma_k^2)^{-\alpha-1}\exp\Bigg[-\frac{\eta}{\sigma_k^2}\Bigg] \\ &\propto (\sigma_k^2)^{-\alpha-1}\exp\Bigg[ -\frac{\eta}{\sigma_k^2}\Bigg] \end{aligned} \tag{54a}$$

Por su parte, aplicando 6 y 8:

$$
\begin{aligned}
p(\sigma_k^2 \mid resto) &\propto \prod_{j=1}^{n_k}\left(\prod_{i=1}^{n_{jk}}\frac{1}{\sqrt{2\pi\sigma_k^2}}\left[\exp\left(-\frac{1}{2\sigma_k^2}\Big(y_{ijk}-\theta_{jk}\Big)^2\right)\right]\right) * (\sigma_k^2)^{-\alpha-1}\exp\left[-\frac{\eta}{\sigma_k^2}\right] \\
&= \prod_{j=1}^{n_k}\left((2\pi\sigma_k^2)^{-(n_{jk}/2)}\exp\left[-\frac{1}{2\sigma_k^2}\sum_{i=1}^{n_{jk}}\left(y_{ijk}-\theta_{jk}\right)^2\right]\right)(\sigma_k^2)^{-\alpha-1}\exp\left[-\frac{\eta}{\sigma_k^2}\right] \\
&\propto (\sigma_k^2)^{-\frac{1}{2}(\sum_{j=1}^{n_k} n_{jk})}\exp\left[-\frac{1}{2\sigma_k^2}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\left(y_{ijk}-\theta_{jk}\right)^2\right](\sigma_k^2)^{-\alpha-1}\exp\left[-\frac{\eta}{\sigma_k^2}\right] \\
&\propto (\sigma_k^2)^{-(\alpha+\frac{1}{2}\sum_{j=1}^{n_k} n_{jk}+1)}\exp\left(-\frac{1}{\sigma_k^2}\left[\eta+\frac{1}{2}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\left(y_{ijk}-\theta_{jk}\right)^2\right]\right) \qquad (55a)
\end{aligned}
$$

que tiene la forma de una distribución Gamma inversa, siendo

$$
C_{jk} = \alpha + \frac{1}{2}\sum_{j=1}^{n_k} n_{jk} \qquad (56)
$$

y

$$
D_{jk} = \eta + \frac{1}{2}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\left(y_{ijk}-\theta_{jk}\right)^2 \qquad (57)
$$

Implica que la ecuación (55a) se escriba de manera sucinta así:

$$
p(\sigma_k^2 \mid \theta_{jk}, y_{ijk}, \alpha, \eta) \propto (\sigma_k^2)^{-C_{jk}-1}\exp\left[-\frac{1}{\sigma_k^2}D_{jk}\right] \qquad (58)
$$

---

Dadas las definiciones de las ecuaciones (7) y (9), se desarrolla la distribución de $\mu_k$:

$$
\begin{aligned}
p(\mu_k \mid resto) &\propto \frac{1}{\sqrt{2\pi\kappa^2}}\exp\left[-\frac{1}{2\kappa^2}(\mu_k-\gamma)^2\right]\prod_{j=1}^{n_k}\frac{1}{\sqrt{2\pi\tau_k^2}}\exp\left[-\frac{1}{2\tau_k^2}(\theta_{jk}-\mu_k)^2\right] \\
&\propto \exp\left[-\frac{1}{2\kappa^2}(\mu_k^2-2\mu_k\gamma)\right]\exp\left[-\frac{1}{2\tau_k^2}\Big(\sum_{j=1}^{n_k}\mu_k^2-2\sum_{j=1}^{n_k}\mu_k\theta_{jk}\Big)\right] \\
&= \exp\left[-\frac{1}{2\kappa^2}(\mu_k^2-2\mu_k\gamma_k)\right]\exp\left[-\frac{1}{2\tau_k^2}\Big(n_k\mu_k^2-2\mu_k\sum_{j=1}^{n_k}\theta_{jk}\Big)\right] \\
&\propto \exp\left[-\frac{1}{2}\Big(\mu_k^2\Big[\frac{1}{\kappa^2}+\frac{n_k}{\tau_k^2}\Big]-2\mu_k\Big[\frac{\gamma}{\kappa^2}+\frac{\sum_{j=1}^{n_k}\theta_{jk}}{\tau_k^2}\Big]\Big)\right] \qquad (59a)
\end{aligned}
$$

Por tanto, se obtiene un resultado similar al de los factores a y b de la ecuación (50a). Con:

$$\tilde{A} = \left(\frac{1}{\kappa^2} + \frac{n_k}{\tau_k^2}\right) = \frac{\tau_k^2 + n_k\kappa^2}{\kappa^2\tau_k^2}, \tag{60a}$$

$$\tilde{B} = \left(\frac{1}{\kappa^2}\gamma + \frac{\sum_{i=1}^{n_{jk}}\theta_{jk}}{\tau_k^2}\right) = \left(\frac{1}{\kappa^2}\gamma + \frac{n_{jk}}{\tau_k^2}\bar{\theta}_{jk}\right), \tag{60b}$$

Sea $E = \tilde{B}/\tilde{A}$ y $F^2 = 1/\tilde{A}$

$$(\mu_k \mid \gamma, \theta_{jk}, \kappa^2, \tau_k^2) \overset{\text{iid}}{\sim} N(E, F^2) \tag{61}$$

---

Y dadas las definiciones de las ecuaciones (45a) y (10), se desarrolla la distribución de $\tau_k^2$:

$$\begin{aligned}(\tau_k^2 \mid resto) &= \prod_{j=1}^{n_k} \frac{1}{\sqrt{2\pi\tau_k^2}} \exp\left[-\frac{1}{2\tau_k^2}\left[(\theta_{jk} - \mu_k)^2\right]\right] (\tau_k^2)^{-(\lambda+1)} \exp\left[-\frac{\xi}{\tau_k^2}\right] \\ &\propto (\tau_k^2)^{-(\lambda+\frac{n_k}{2}+1)} \exp\left[-\frac{1}{\tau_k^2}\left(\frac{1}{2}\sum_{j=1}^{n_k}(\theta_{jk} - \mu_k)^2 + \xi\right)\right]\end{aligned} \tag{62a}$$

Sean

$$G = (\lambda + \frac{n_k}{2}) \tag{63a}$$

$$H = \left(\frac{1}{2}\sum_{j=1}^{n_k}(\theta_{jk} - \mu_k)^2 + \xi\right) \tag{63b}$$

Entonces,

$$(\tau_k^2 \mid \lambda, \xi, \theta_{jk}, \mu_k) \overset{\text{iid}}{\sim} IGamma(G, H) \tag{64}$$

---

Dada la definición de (11):

Para $\alpha$:

$$\begin{aligned}p(\alpha \mid resto) = p(\alpha)p(\sigma_k^2 \mid resto) &\propto \exp(-\alpha a_0) \prod_{k=1}^{M} IGamma(\sigma_k^2 \mid \alpha, \eta) \\ &= \exp(-\alpha a_0) \prod_{k=1}^{M} \left(\frac{\eta^\alpha}{\Gamma(\alpha)}(\sigma_k^2)^{-\alpha-1} exp\Big[-\frac{\eta}{\sigma_k^2}\Big]\right) \\ &\propto \exp(-\alpha a_0)\left[\frac{\eta^{M\alpha}}{\Gamma(\alpha)^M}\prod_{k=1}^{M}(\sigma_k^2)^{-(\alpha+1)}\right]\end{aligned} \tag{65a}$$

---

Dada la definición de (12):

$$
\begin{aligned}
p(\eta \mid resto) &\propto \prod_{k=1}^{M} IGamma(\sigma_k^2 \mid \alpha, \eta) Gamma(\eta \mid b_0, c_0) \\
&= \prod_{k=1}^{M} \left( \frac{\eta^{\alpha}}{\Gamma(\alpha)} (\sigma_k^2)^{-\alpha-1} exp\Big[ -\frac{\eta}{\sigma_k^2} \Big] \right) \left[ \frac{c_0^{b_0}}{\Gamma(b_0)} \eta^{(b_0-1)} exp(-c_0\eta) \right] \\
&\propto \eta^{M\alpha} exp\Big[ -\eta \sum_{k=1}^{M} \frac{1}{\sigma_k^2} \Big] \eta^{(b_0-1)} exp(-c_0\eta) \\
&= \eta^{b_0+M\alpha-1} exp\Big[ -\Big( c_0 + \sum_{k=1}^{M} \frac{1}{\sigma_k^2} \Big) \eta \Big] \qquad (66a)
\end{aligned}
$$

Sea $K = b_0 + M\alpha$ y $L = c_0 + \sum_{k=1}^{M} \frac{1}{\sigma_k^2}$, entonces,

$$p(\eta \mid \alpha, \sigma_k^2) \sim Gamma(K, L)$$

---

Dada la definición de (13):

Para $\gamma$:

$$
\begin{aligned}
p(\gamma \mid resto) &\propto \prod_{k=1}^{M} \Big(N(\mu_k \mid \gamma, \kappa^2)\Big) N(\gamma \mid g_0, h_0^2) \\
&= \prod_{k=1}^{M} \left(\frac{1}{\sqrt{2\pi\kappa^2}} exp\Big[-\frac{1}{2\kappa^2}(\mu_k - \gamma)^2\Big]\right)\left(\frac{1}{\sqrt{2\pi h_0^2}} exp\Big[-\frac{1}{2h_0^2}(\gamma - g_0)^2\Big]\right) \\
&\propto \prod_{k=1}^{M} \left(exp\Big[-\frac{1}{2\kappa^2}(\mu_k - \gamma)^2\Big]\right)\left(exp\Big[-\frac{1}{2h_0^2}(\gamma - g_0)^2\Big]\right) \\
&= \prod_{k=1}^{M} \left(exp\Big[-\frac{1}{2\kappa^2}(\mu_k^2 - 2\mu_k\gamma + \gamma^2)\Big]\right)\left(exp\Big[-\frac{1}{2h_0^2}(\gamma^2 - 2\gamma g_0 + g_0^2)\Big]\right) \\
&\propto \left(exp\Big[-\frac{1}{2\kappa^2}(-2\gamma \sum_{k=1}^{M}(\mu_k) + M\gamma^2)\Big]\right)\left(exp\Big[-\frac{1}{2h_0^2}(\gamma^2 - 2\gamma g_0)\Big]\right) \\
&= exp\left[-\frac{1}{2}\left[\gamma^2\left(\frac{M}{\kappa^2} + \frac{1}{h_0^2}\right) - 2\gamma\left(\frac{\sum_{k=1}^{M}(\mu_k)}{\kappa^2} + \frac{g_0}{h_0^2}\right)\right]\right] \\
&= exp\left[-\frac{1}{2}\left[\gamma^2\left(\frac{Mh_0^2 + \kappa^2}{\kappa^2 h_0^2}\right) - 2\gamma\left(\frac{1}{h_0^2}g_0 + \frac{\sum_{k=1}^{M}(\mu_k)}{\kappa^2}\right)\right]\right] \qquad (67a)
\end{aligned}
$$

Si $\tilde{A} = \frac{Mh_0^2+\kappa^2}{\kappa^2 h_0^2}$ y $\tilde{B} = \frac{1}{h_0^2}g_0 + \frac{\sum_{k=1}^{M}(\mu_k)}{\kappa^2} = \frac{1}{h_0^2}g_0 + \frac{M}{\kappa^2}\bar{\mu}_k$ y sea $W = \tilde{B}/\tilde{A}$ y $X^2 = 1/\tilde{A}$,

entonces $p(\gamma \mid \mu_k, \kappa^2) \sim N(W, X^2)$

---

Y para $\kappa^2$ (ver definición de (14)):

$$
\begin{aligned}
p(\kappa^2 \mid resto) &\propto \prod_{k=1}^{M} \Big(N(\mu_k \mid \gamma, \kappa^2\Big) IGamma(\kappa^2 \mid u_0, v_0) \\
&\propto \prod_{k=1}^{M} \left(\frac{1}{\sqrt{2\pi\kappa^2}} exp\Big[-\frac{1}{2\kappa^2}(\mu_k - \gamma)^2\Big]\right)\left((\kappa^2)^{-u_0-1} exp\big[-\frac{v_0}{\kappa^2}\big]\right) \\
&\propto (\kappa^2)^{-(u_0+\frac{M}{2}+1)} exp\left[-\frac{1}{\kappa^2}\left(\frac{1}{2}\sum_{k=1}^{M}(\mu_k - \gamma)^2 + v_0\right)\right] \qquad (68a)
\end{aligned}
$$

Sean $Y = u_0 + \frac{M}{2}$ y $Z = v_0 + \frac{1}{2}\sum_{k=1}^{M}(\mu_k - \gamma)^2$, entonces $p(\kappa^2 \mid \mu_k, \gamma) \sim IGamma(Y, Z)$

---

Dada la definición de (15):

Para $\lambda$:

$$p(\lambda \mid resto) \propto \prod_{k=1}^{M} \Big( IGamma(\tau_k^2 \mid \lambda, \xi) \Big) exp(-\lambda d_0) \tag{69a}$$

$$= \prod_{k=1}^{M} \left[ \frac{\xi^\lambda}{\Gamma(\lambda)} (\tau_k^2)^{-(\lambda+1)} exp\Big( -\frac{\xi}{\tau_k^2} \Big) \right] exp(-\lambda d_0) \tag{69b}$$

$$\propto \frac{[\xi^\lambda]^M}{\big[\Gamma(\lambda)\big]^M} \Big( \prod_{k=1}^{M} \tau_k^2 \Big)^{-(\lambda+1)} exp(-\lambda d_0) \tag{69c}$$

---

Y para $\xi$ (ver ecuación (16)):

$$p(\xi \mid resto) \propto \prod_{k=1}^{M} IGamma(\tau_k^2 \mid \lambda, \xi) Gamma(\xi \mid e_0, f_0) \tag{70a}$$

$$= \prod_{k=1}^{M} \left[ \frac{\xi^\lambda}{\Gamma(\lambda)} (\tau_k^2)^{-(\lambda+1)} exp\Big( -\frac{\xi}{\tau_k^2} \Big) \right] \frac{f_0^{e_0}}{\Gamma(e_0)} \xi^{(e_0-1)} exp(-f_0 \xi) \tag{70b}$$

$$\propto \xi^{\lambda M} exp\left( -\xi \sum_{k=1}^{M} \frac{1}{\tau_k^2} \right) \xi^{(e_0-1)} exp(-f_0 \xi) \tag{70c}$$

$$= \xi^{e_0+\lambda M-1} exp\Big[ -\Big( f_0 + \sum_{k=1}^{M} \frac{1}{\tau_k^2} \Big) \xi \Big] \tag{70d}$$

Sea $Q = e_0 + M\lambda$ y $R = f_0 + \sum_{k=1}^{M} \frac{1}{\tau_k^2}$, entonces $p(\xi \mid \lambda, \tau_k^2) \sim Gamma(Q, R)$

| **Niv.** | **Par.** | **Distribución** | **Previa** | **Posterior** | **Siendo** |
|---|---|---|---|---|---|
| I | $\theta_{jk}$ | $\{Y_{ijk} \mid \theta_{jk}, \sigma_k^2\} \overset{\text{ind}}{\sim} N(\theta_{jk}, \sigma_k^2)$ | $\{\theta_{jk} \mid \mu_k, \tau_k^2\} \overset{\text{ind}}{\sim} N(\mu_k, \tau_k^2)$ | $p(\theta_{jk} \mid \sigma_k^2, y_{ijk}, \mu_k, \tau_k^2) \sim N(A_{jk}, B_{jk}^2)$ | $A_{jk} = \left[\frac{1}{\tau_k^2}\mu_k + \frac{n_{jk}}{\sigma_k^2}\bar{y}_{jk}\right]\left[\frac{\tau_k^2\sigma_k^2}{\sigma_k^2+n_{jk}\tau_k^2}\right]$ y $B_{jk}^2 = \frac{\tau_k^2\sigma_k^2}{\sigma_k^2+\tau_k^2 n_{jk}}$ |
| | $\sigma_k^2$ | | $\{\sigma_k^2 \mid \alpha, \eta\} \overset{\text{ind}}{\sim} IGamma(\alpha, \eta)$ | $p(\sigma_k^2 \mid \theta_{jk}, y_{ijk}, \alpha, \eta) \sim IGamma(C_{jk}, D_{jk})$ | $C_{jk} = \alpha + \frac{1}{2}\sum_{j=1}^{n_k} n_{jk}$ y $D_{jk} = \eta + \frac{1}{2}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\left(y_{ijk} - \theta_{jk}\right)^2$ |
| II | $\mu_k$ | $\{\theta_{jk} \mid \mu_k, \tau_k^2\} \overset{\text{ind}}{\sim} N(\mu_k, \tau_k^2)$ | $\{\mu_k \mid \gamma, \kappa^2\} \overset{\text{iid}}{\sim} N(\gamma, \kappa^2)$ | $p(\mu_k \mid \gamma, \theta_{jk}, \kappa^2, \tau_k^2) \sim N(E, F^2)$ | $E = \left[\frac{\gamma}{\kappa^2} + \frac{\sum_{j=1}^{n_k}\theta_{jk}}{\tau_k^2}\right]\left[\frac{\kappa^2\tau_k^2}{\tau_k^2+n_k\kappa^2}\right]$ y $F^2 = \frac{\kappa^2\tau_k^2}{\tau_k^2+n_k\kappa^2}$ |
| | $\tau_k^2$ | | $\tau_k^2 \overset{\text{iid}}{\sim} IGamma(\lambda, \xi)$ | $p(\tau_k^2 \mid \lambda, \xi, \theta_{jk}, \mu_k) \sim IGamma(G, H)$ | $G = (\lambda + \frac{n_k}{2})$ y $H = \frac{1}{2}\sum_{j=1}^{n_k}(\theta_{jk} - \mu_k)^2 + \xi$ |
| II | $\alpha$ | $\{\sigma_k^2 \mid \alpha, \eta\} \overset{\text{ind}}{\sim} IGamma(\alpha, \eta)$ | $p(\alpha) \propto e^{-\alpha a_0}$ | $p(\alpha \mid \sigma_k^2, \eta) \propto \left[\frac{\eta^{M\alpha}}{\Gamma(\alpha)^M}\prod_{k=1}^{M}(\sigma_k^2)^{-(\alpha+1)}\right]e^{(-\alpha a_0)}$ | – |
| | $\eta$ | | $p(\eta) \propto Gamma(b_{0k}, c_{0k})$ | $p(\eta \mid \alpha, \sigma_k^2) \sim Gamma(K, L)$ | con $K = b_0 + M\alpha$ y $L = c_0 + \sum_{k=1}^{M}\frac{1}{\sigma_k^2}$ |
| III | $\gamma$ | $\{\mu_k \mid \gamma, \kappa^2\} \overset{\text{iid}}{\sim} N(\gamma, \kappa^2)$ | $\gamma \sim N(g_0, h_0^2)$ | $p(\gamma \mid \kappa, \mu_k) \sim N(W, X^2)$ | $W = \left[\frac{1}{h_0^2}g_0 + \frac{\sum_{k=1}^{M}\mu_k}{\kappa^2}\right]\left[\frac{\kappa^2 h_0^2}{Mh_0^2+\kappa^2}\right]$ y $X^2 = \frac{\kappa^2 h_0^2}{Mh_0^2+\kappa^2}$ |
| | $\kappa^2$ | | $\kappa^2 \sim IGamma(u_0, v_0)$ | $p(\kappa^2 \mid \mu_k, \gamma) \sim IGamma(Y, Z)$ | $Y = u_0 + \frac{M}{2}$ y $Z = v_0 + \frac{1}{2}\left[\sum_{k=1}^{M}(\mu_k - \gamma)^2\right]$ |
| III | $\lambda$ | $\{\tau_k^2 \mid \lambda, \xi\} \overset{\text{iid}}{\sim} IGamma(\lambda, \xi)$ | $\lambda \propto e^{-\lambda d_0}$ | $p(\lambda \mid \tau_k^2, \xi) \sim \left[\frac{\xi^{M\lambda}}{\Gamma(\lambda)^M}\left[\prod_{k=1}^{M}\tau_k^2\right]^{-(\lambda+1)}\right]e^{-\lambda d_0}$ | – |
| | $\xi$ | | $\xi \sim Gamma(e_0, f_0)$ | $p(\xi \mid \lambda, \tau_k^2) \sim Gamma(Q, R)$ | $Q = e_0 + M\lambda$ y $R = f_0 + \sum_{k=1}^{M}\frac{1}{\tau_k^2}$ |

Cuadro 14: Resumen del modelo jerárquico de tres niveles.

## Anexo B. Inferencia de las distribuciones condicionales conjuntas del tercer modelo: jerárquico de dos niveles

Desarrollo de la verosimilitud a **nivel de aula** partir de la definición dada en la ecuación (20) para una institución educativa cualquiera k:

$$
\begin{aligned}
p(y_{ijk} \mid \boldsymbol{\beta}_{jk}, \sigma_k^2, \boldsymbol{X}_{ijk}) &= \prod_{i=1}^{n_{jk}} \Big[(2\pi)^{-p/2} \mid \sigma_k^2 \mid^{-1/2} \exp\Big( -\frac{1}{2}(y_{ijk} - \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk})^T(\sigma_k^2)^{-1}(y_{ijk} - \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk})\Big)\Big] \\
&= (2\pi)^{-n_{jk}p/2} \mid \sigma_k^2 \mid^{-n_{jk}p/2} \prod_{i=1}^{n_{jk}} (\exp\Big[ -\frac{1}{2}(y_{ijk} - \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk})^T(\sigma_k^2)^{-1}(y_{ijk} - \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk})\Big] \\
&\propto \prod_{i=1}^{n_{jk}} (\exp\Big[ -\frac{1}{2}(y_{ijk} - \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk})^T(\sigma_k^2)^{-1}(y_{ijk} - \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk})\Big] \qquad (71a)
\end{aligned}
$$

Se desarrolla el modelo de la previa de $\boldsymbol{\beta}_{jk}$

Por definición de la ecuación (22):

$$
\begin{aligned}
p(\boldsymbol{\beta_{jk}} \mid \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) &\propto (2\pi)^{-p/2} \mid \boldsymbol{\Sigma_k} \mid^{-1/2} \exp\Big[ -\frac{1}{2}(\boldsymbol{\beta}_{jk} - \boldsymbol{\psi}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\beta}_{jk} - \boldsymbol{\psi}_k)\Big] \\
&= (2\pi)^{-p/2} \mid \boldsymbol{\Sigma}_k \mid^{-1/2} \exp\Big( -\frac{1}{2}\boldsymbol{\beta}_{jk}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{jk} + \boldsymbol{\beta}_{jk}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\psi}_k - \frac{1}{2}\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\psi}_k\Big) \\
&\propto \exp\Big[ -\frac{1}{2}\Big(\boldsymbol{\beta}_{jk}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{jk} - 2\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{jk}\Big)\Big] \qquad (72a)
\end{aligned}
$$

---

Posteriores:

De las ecuaciones (71) y (72):

$$
\begin{aligned}
&p(\boldsymbol{\beta}_{jk} \mid resto) \propto p(\boldsymbol{y}_{ijk} \mid \boldsymbol{\beta}_{jk}, \sigma_k^2, \boldsymbol{X}_{ijk}) p(\boldsymbol{\beta}_{jk} \mid \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) \\
&\propto \prod_{i=1}^{n_{jk}} \left( \exp\Big[ -\frac{1}{2}\sum_{i=1}^{n_{jk}} (y_{ijk} - \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk})^T(\sigma_k^2)^{-1}(y_{ijk} - \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk})\Big] \right) \exp\Big[ -\frac{1}{2}\Big(\boldsymbol{\beta}_{jk}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{jk} - 2\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{jk}\Big)\Big] \\
&= \prod_{i=1}^{n_{jk}} \left[ \exp\left( -\frac{1}{2}\Big[\boldsymbol{\beta}_{jk}^T(\boldsymbol{X}_{ijk}\sigma_k^{-2}\boldsymbol{X}_{ijk}^T)\boldsymbol{\beta}_{jk} - 2\sigma_k^{-2} y_{ijk}\boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk}\Big] \right) \right] \exp\Big[ -\frac{1}{2}\Big(\boldsymbol{\beta}_{jk}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{jk} - 2\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{jk}\Big)\Big] \\
&= \exp\left( -\frac{1}{2}\Big[\boldsymbol{\beta}_{jk}^T(\sigma_k^{-2}\sum_{i=1}^{n_{jk}} \boldsymbol{X}_{ijk}\boldsymbol{X}_{ijk}^T)\boldsymbol{\beta}_{jk} - 2\sigma_k^{-2}\sum_{i=1}^{n_{jk}} \Big(y_{ijk}\boldsymbol{X}_{ijk}^T\Big)\boldsymbol{\beta}_{jk}\Big] \right) \exp\Big[ -\frac{1}{2}\Big(\boldsymbol{\beta}_{jk}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{jk} - 2\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{jk}\Big)\Big] \\
&\propto \exp\left( -\frac{1}{2}\Big[\boldsymbol{\beta}_{jk}^T\Big(\boldsymbol{\Sigma}_k^{-1} + \sigma_k^{-2}\sum_{i=1}^{n_{jk}} \boldsymbol{X}_{ijk}\boldsymbol{X}_{ijk}^T\Big)\boldsymbol{\beta}_{jk} - 2\Big(\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1} + \sigma_k^{-2}\sum_{i=1}^{n_{jk}} \Big(y_{ijk}\boldsymbol{X}_{ijk}^T\Big)\Big)\boldsymbol{\beta}_{jk}\Big] \right) \qquad (73a)
\end{aligned}
$$

Con

$$\boldsymbol{V}_{jk} = \Big(\boldsymbol{\Sigma}_k^{-1} + \sigma_k^{-2}\sum_{i=1}^{n_{jk}} \boldsymbol{X}_{ijk}\boldsymbol{X}_{ijk}^T\Big)^{-1} \tag{74}$$

y

$$\boldsymbol{m}_{jk} = \boldsymbol{V}_{jk}\Big(\boldsymbol{\Sigma_k^{-1}}\boldsymbol{\psi}_k + \sigma_k^{-2}\sum_{i=1}^{n_{jk}}(y_{ijk}\boldsymbol{X}_{ijk}^T)\Big) \tag{75}$$

$$p(\boldsymbol{\beta}_{jk} \mid \boldsymbol{y}_{ijk}, \boldsymbol{\Sigma}_k, \boldsymbol{\psi}_k, \sigma_k^2, \boldsymbol{X}_{ijk}) \sim N(\boldsymbol{m}_{jk}, \boldsymbol{V}_{jk}) \tag{76}$$

---

Desarrollo de la verosimilitud a **nivel de establecimiento educativo** a partir de la definición dada en la ecuación (21)

$$\begin{aligned} p(y_{ijk} \mid \boldsymbol{\beta}_k, \sigma_k^2, \boldsymbol{X}_{ijk}) &= \Big[(2\pi)^{-p/2} \mid \sigma_k^2 \mid^{-1/2} \exp\Big(-\frac{1}{2}(y_{ijk} - \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_k)^T(\sigma_k^2)^{-1}(y_{ijk} - \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_k)\Big)\Big] \\ &\propto (\exp\Big[-\frac{1}{2}(\sigma_k^2)^{-1}(y_{ijk} - \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_k)^T(y_{ijk} - \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_k)\Big] \end{aligned} \tag{77a}$$

Se desarrolla el modelo de la previa de $\boldsymbol{\beta}_k$

Por definición de la ecuación (23) para una institución educativa cualquiera k:

$$\begin{aligned} p(\boldsymbol{\beta_k} \mid \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) &\propto (2\pi)^{-p/2} \mid \boldsymbol{\Sigma_k} \mid^{-1/2} \exp\Big[-\frac{1}{2}(\boldsymbol{\beta}_k - \boldsymbol{\psi}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\beta}_k - \boldsymbol{\psi}_k)\Big] \\ &= (2\pi)^{-p/2} \mid \boldsymbol{\Sigma}_k \mid^{-1/2} \exp\Big(-\frac{1}{2}\boldsymbol{\beta}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k + \boldsymbol{\beta}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\psi}_k - \frac{1}{2}\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\psi}_k\Big) \\ &\propto \exp\Big[-\frac{1}{2}\Big(\boldsymbol{\beta}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k - 2\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}k\Big)\Big] \end{aligned} \tag{78a}$$

---

Posteriores:

De las ecuaciones (77) y (78):

$$
\begin{aligned}
p(\boldsymbol{\beta}_k \mid resto) &\propto \prod_{j=1}^{n_k}\prod_{i=1}^{n_{jk}} p(\boldsymbol{y}_{ijk} \mid \boldsymbol{\beta}_k, \sigma_k^2, \boldsymbol{X}_{ijk}) p(\boldsymbol{\beta}_k \mid \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) \\
&\propto \prod_{j=1}^{n_k}\prod_{i=1}^{n_{jk}} \exp\Big[-\frac{1}{2}(\sigma_k^2)^{-1}(y_{ijk}-\boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_k)^T(y_{ijk}-\boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_k)\Big]\exp\Big[-\frac{1}{2}\Big(\boldsymbol{\beta}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k - 2\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k\Big)\Big] \\
&= \prod_{j=1}^{n_k}\prod_{i=1}^{n_{jk}} \exp\Bigg(-\frac{1}{2}\Big[\boldsymbol{\beta}_k^T(\boldsymbol{X}_{ijk}\sigma_k^{-2}\boldsymbol{X}_{ijk}^T)\boldsymbol{\beta}_k - 2\sigma_k^{-2}y_{ijk}\boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_k\Big]\Bigg)\exp\Big[-\frac{1}{2}\Big(\boldsymbol{\beta}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k - 2\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k\Big)\Big] \\
&= \prod_{j=1}^{n_k}\prod_{i=1}^{n_{jk}} \exp\Bigg(-\frac{1}{2}\Big[\boldsymbol{\beta}_k^T(\sigma_k^{-2}\boldsymbol{X}_{ijk}\boldsymbol{X}_{ijk}^T)\boldsymbol{\beta}_k - 2\sigma_k^{-2}\Big(y_{ijk}\boldsymbol{X}_{ijk}^T\Big)\boldsymbol{\beta}_k\Big]\Bigg)\exp\Big[-\frac{1}{2}\Big(\boldsymbol{\beta}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k - 2\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k\Big)\Big] \\
&= \exp\Bigg(-\frac{1}{2}\Big[\boldsymbol{\beta}_k^T\Big(\boldsymbol{\Sigma}_k^{-1} + \sigma_k^{-2}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}(\boldsymbol{X}_{ijk}\boldsymbol{X}_{ijk}^T)\Big)\boldsymbol{\beta}_k - 2\Big(\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1} + \sigma_k^{-2}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\Big(y_{ijk}\boldsymbol{X}_{ijk}^T\Big)\Big)\boldsymbol{\beta}_k\Big]\Bigg)
\end{aligned}
\tag{79a}
$$

Con

$$
\boldsymbol{V}_k = \Big(\boldsymbol{\Sigma}_k^{-1} + \sigma_k^{-2}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}(\boldsymbol{X}_{ijk}\boldsymbol{X}_{ijk}^T)\Big)^{-1} \tag{80}
$$

y

$$
\boldsymbol{m}_k = \boldsymbol{V}_k\Big(\boldsymbol{\Sigma}_{\boldsymbol{k}}^{-\mathbf{1}}\boldsymbol{\psi}_k + \sigma_k^{-2}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}(y_{ijk}\boldsymbol{X}_{ijk}^T)\Big) \tag{81}
$$

$$
p(\boldsymbol{\beta}_k \mid \boldsymbol{y}_{ijk}, \boldsymbol{\Sigma}_k, \boldsymbol{\psi}_k, \sigma_k^2, \boldsymbol{X}_{ijk}) \sim N(\boldsymbol{m}_k, \boldsymbol{V}_k) \tag{82}
$$

---

A partir de la ecuación (25):

$p(\boldsymbol{\Sigma}_k) = \mid \boldsymbol{\Sigma}_k \mid^{-(\nu_0+p+1)/2} \exp\Big[-\frac{1}{2}tr(\mathbf{S}_0\boldsymbol{\Sigma}_k^{-1})\Big]$

Entonces, sobre el modelo **a nivel de aula**:

$$
\begin{aligned}
p(\boldsymbol{\Sigma}_k \mid resto) &\propto p(\boldsymbol{\Sigma}_k)\prod_{j=1}^{n_k}(p(\boldsymbol{\beta}_{jk} \mid \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)) \\
&\propto \mid \boldsymbol{\Sigma}_k \mid^{-(\nu_0+p+1)/2} \exp\Big[-\frac{1}{2}tr(\mathbf{S}_0\boldsymbol{\Sigma}_k^{-1})\Big]\prod_{j=1}^{n_k} \mid \boldsymbol{\Sigma}_k \mid^{-1/2} \exp\Big[-\frac{1}{2}(\boldsymbol{\beta}_{jk}-\boldsymbol{\psi}_k)^T\Sigma_k^{-1}(\boldsymbol{\beta}_{jk}-\boldsymbol{\psi}_k)\Big] \\
&= \mid \boldsymbol{\Sigma}_k \mid^{-(\nu_0+p+1)/2} \mid \boldsymbol{\Sigma}_k \mid^{-n_k/2} \exp\Big[-\frac{1}{2}tr(\mathbf{S}_0\boldsymbol{\Sigma}_k^{-1})\Big]\exp\Big[-\frac{1}{2}\sum_{j=1}^{n_k}\Big((\boldsymbol{\beta}_{jk}-\boldsymbol{\psi}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\beta}_{jk}-\boldsymbol{\psi}_k)\Big)\Big],
\end{aligned}
$$

Basado en que la traza de un numero es el mismo número, que $tr(a+b) = tr(a) + tr(b)$ y que $tr(a*b) = tr(b*a)$, se pueden realizar las siguientes manipulaciones algebraicas:

$$
\begin{aligned}
p(\boldsymbol{\Sigma}_k \mid resto) &\propto \mid \boldsymbol{\Sigma}_k \mid^{-(\nu_0+n_k+p+1)/2} \exp\Big[-\frac{1}{2}tr(\mathbf{S}_0\boldsymbol{\Sigma}_k^{-1})\Big]\exp\Big[-\frac{1}{2}tr\Big(\sum_{j=1}^{n_k}\Big((\boldsymbol{\beta}_{jk}-\boldsymbol{\psi}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\beta}_{jk}-\boldsymbol{\psi}_k)\Big)\Big)\Big] \\
&= \mid \boldsymbol{\Sigma}_k \mid^{-(\nu_0+n_k+p+1)/2} \exp\Big[-\frac{1}{2}tr(\mathbf{S}_0\boldsymbol{\Sigma}_k^{-1})\Big]\exp\Big[-\frac{1}{2}\sum_{j=1}^{n_k}tr\Big((\boldsymbol{\beta}_{jk}-\boldsymbol{\psi}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\beta}_{jk}-\boldsymbol{\psi}_k)\Big)\Big] \\
&= \mid \boldsymbol{\Sigma}_k \mid^{-(\nu_0+n_k+p+1)/2} \exp\Big[-\frac{1}{2}tr(\mathbf{S}_0\boldsymbol{\Sigma}_k^{-1})\Big]\exp\Big[-\frac{1}{2}\sum_{j=1}^{n_k}tr\Big((\boldsymbol{\beta}_{jk}-\boldsymbol{\psi}_k)(\boldsymbol{\beta}_{jk}-\boldsymbol{\psi}_k)^T\boldsymbol{\Sigma}_k^{-1}\Big)\Big] \\
&= \mid \boldsymbol{\Sigma}_k \mid^{-(\nu_0+n_k+p+1)/2} \exp\Big[-\frac{1}{2}tr(\mathbf{S}_0\boldsymbol{\Sigma}_k^{-1})\Big]\exp\Big[-\frac{1}{2}tr\Big(\sum_{j=1}^{n_k}\Big((\boldsymbol{\beta}_{jk}-\boldsymbol{\psi}_k)(\boldsymbol{\beta}_{jk}-\boldsymbol{\psi}_k)^T\boldsymbol{\Sigma}_k^{-1}\Big)\Big)\Big]
\end{aligned}
$$

Este resultado lleva a:

$$p(\boldsymbol{\Sigma}_k \mid \boldsymbol{\beta}_{jk}, \boldsymbol{\Psi}_k) \sim IWishart\Big(\nu_k, S_k\Big) \tag{85}$$

con $\nu_k = \nu_0 + n_k$ y $\mathbf{S_k} = \mathbf{S}_0 + \sum_{j=1}^{n_k}\Big((\boldsymbol{\beta}_{jk}-\boldsymbol{\psi}_k)(\boldsymbol{\beta}_{jk}-\boldsymbol{\psi}_k)^T\Big)$

Y sobre el modelo a **nivel de establecimiento educativo**:

$$
\begin{aligned}
p(\boldsymbol{\Sigma}_k \mid resto) &\propto p(\boldsymbol{\Sigma}_k)(p(\boldsymbol{\beta}_k \mid \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)) \\
&\propto \mid \boldsymbol{\Sigma}_k \mid^{-(\nu_0+p+1)/2} \exp\Big[-\frac{1}{2}tr(\mathbf{S}_0\boldsymbol{\Sigma}_k^{-1})\Big] \mid \boldsymbol{\Sigma}_k \mid^{-1/2} \exp\Big[-\frac{1}{2}(\boldsymbol{\beta}_k-\boldsymbol{\psi}_k)^T\Sigma_k^{-1}(\boldsymbol{\beta}_k-\boldsymbol{\psi}_k)\Big] \\
&= \mid \boldsymbol{\Sigma}_k \mid^{-(\nu_0+p+1)/2} \mid \boldsymbol{\Sigma}_k \mid^{-1/2} \exp\Big[-\frac{1}{2}tr(\mathbf{S}_0\boldsymbol{\Sigma}_k^{-1})\Big]\exp\Big[-\frac{1}{2}\Big((\boldsymbol{\beta}_k-\boldsymbol{\psi}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\beta}_k-\boldsymbol{\psi}_k)\Big)\Big],
\end{aligned}
$$

Basado en que la traza de un numero es el mismo número:

$$p(\boldsymbol{\Sigma}_k \mid resto) \propto \mid \boldsymbol{\Sigma}_k \mid^{-(\nu_0+p)/2} \exp\Big[-\frac{1}{2}tr(\mathbf{S}_0\boldsymbol{\Sigma}_k^{-1})\Big]\exp\Big[-\frac{1}{2}tr\Big(\Big((\boldsymbol{\beta}_k-\boldsymbol{\psi}_k)(\boldsymbol{\beta}_k-\boldsymbol{\psi}_k)^T\boldsymbol{\Sigma}_k^{-1}\Big)\Big)\Big]$$

Este resultado lleva a:

$$p(\boldsymbol{\Sigma}_k \mid \boldsymbol{\beta}_k, \boldsymbol{\Psi}_k) \sim IWishart\Big(\nu_k, S_k\Big) \tag{88}$$

con $\nu_k = \nu_0 + (p-1)$ y $\mathbf{S_k} = \mathbf{S}_0 + (\boldsymbol{\beta}_k-\boldsymbol{\psi}_k)(\boldsymbol{\beta}_k-\boldsymbol{\psi}_k)^T$

---

Desarrollo de $p(\boldsymbol{\psi}_k)$:

Basados en la ecuación (24)

$p(\boldsymbol{\psi}_k) = (2\pi\boldsymbol{\Sigma})^{-p/2}\exp\Big[-\frac{1}{2}(\boldsymbol{\psi}_k-\mu)^T\boldsymbol{\Lambda}^{-1}(\boldsymbol{\psi}_k-\mu)\Big]$, entonces, a **nivel de aula**:

$$
\begin{aligned}
p(\boldsymbol{\psi}_k \mid resto) &\propto p(\boldsymbol{\psi}_k)\prod_{j=1}^{n_k} p(\boldsymbol{\beta}_{jk} \mid \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) \\
&\propto \exp\Big[-\frac{1}{2}(\boldsymbol{\psi}_k-\mu)^T\boldsymbol{\Lambda}^{-1}(\boldsymbol{\psi}_k-\mu)\Big]\prod_{j=1}^{n_k}\exp\Big[-\frac{1}{2}\Big(\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\psi}_k - 2\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{jk}\Big)\Big] \\
&\propto \exp\Big[-\frac{1}{2}(\boldsymbol{\psi}_k^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\psi}_k - 2\mu^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\psi}_k)\Big]\exp\Big[-\frac{1}{2}\sum_{j=1}^{n_k}\Big(\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\psi}_k - 2\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{jk}\Big)\Big] \\
&= \exp\Big[-\frac{1}{2}(\boldsymbol{\psi}_k^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\psi}_k) + \mu^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\psi}_k\Big]\exp\Big[-\frac{1}{2}\Big(\sum_{j=1}^{n_k}\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\psi}_k - 2\sum_{j=1}^{n_k}\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{jk}\Big)\Big] \\
&= \exp\Big[-\frac{1}{2}(\boldsymbol{\psi}_k^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\psi}_k) - \frac{1}{2}\sum_{j=1}^{n_k}\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\psi}_k + \psi_k^T(\boldsymbol{\Lambda}^{-1}\mu) + \boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\sum_{j=1}^{n_k}\boldsymbol{\beta}_{jk}\Big] \\
&= \exp\Big[-\frac{1}{2}\Big(\psi_k^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\psi}_k + n_k\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\psi}_k\Big) + \psi_k^T(\boldsymbol{\Lambda}^{-1}\mu) + \boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\sum_{j=1}^{n_k}\boldsymbol{\beta}_{jk}\Big] \\
&= \exp\Bigg[-\frac{1}{2}\Big(\boldsymbol{\psi}_k^T\Big[\boldsymbol{\Lambda}^{-1} + n_k\boldsymbol{\Sigma}_k^{-1}\Big]\boldsymbol{\psi}_k\Big) + \boldsymbol{\psi}_k^T\Big[\boldsymbol{\Sigma}_k^{-1}\sum_{j=1}^{n_k}\boldsymbol{\beta}_{jk} + \boldsymbol{\Lambda}^{-1}\mu\Big]\Bigg] \qquad (89a)
\end{aligned}
$$

Sea $V_k = (\boldsymbol{\Lambda}^{-1} + n_k\boldsymbol{\Sigma}_k^{-1})^{-1}$ y $m_k = V_k(\boldsymbol{\Lambda}^{-1}\mu + \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{\bullet k})$, entonces, $p(\boldsymbol{\psi}_k \mid resto) \sim N(m_k, V_k)$

Y a **nivel de establecimiento educativo**:

$$
\begin{aligned}
p(\boldsymbol{\psi}_k \mid resto) &\propto p(\boldsymbol{\psi}_k)p(\boldsymbol{\beta}_k \mid \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) \\
&\propto \exp\Big[-\frac{1}{2}(\boldsymbol{\psi}_k-\mu)^T\boldsymbol{\Lambda}^{-1}(\boldsymbol{\psi}_k-\mu)\Big]\exp\Big[-\frac{1}{2}\Big(\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\psi}_k - 2\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k\Big)\Big] \\
&\propto \exp\Big[-\frac{1}{2}(\boldsymbol{\psi}_k^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\psi}_k - 2\mu^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\psi}_k)\Big]\exp\Big[-\frac{1}{2}\Big(\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\psi}_k - 2\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{jk}\Big)\Big] \\
&= \exp\Big[-\frac{1}{2}(\boldsymbol{\psi}_k^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\psi}_k) + \mu^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\psi}_k\Big]\exp\Big[-\frac{1}{2}\Big(\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\psi}_k - 2\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k\Big)\Big] \\
&= \exp\Big[-\frac{1}{2}(\boldsymbol{\psi}_k^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\psi}_k) - \frac{1}{2}\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\psi}_k + \psi_k^T(\boldsymbol{\Lambda}^{-1}\mu) + \boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k\Big] \\
&= \exp\Big[-\frac{1}{2}\Big(\psi_k^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\psi}_k + \boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\psi}_k\Big) + \psi_k^T(\boldsymbol{\Lambda}^{-1}\mu) + \boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k\Big] \\
&= \exp\Bigg[-\frac{1}{2}\Big(\boldsymbol{\psi}_k^T\Big[\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Sigma}_k^{-1}\Big]\boldsymbol{\psi}_k\Big) + \boldsymbol{\psi}_k^T\Big[\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k + \boldsymbol{\Lambda}^{-1}\mu\Big]\Bigg] \qquad (90a)
\end{aligned}
$$

Sea $V_k = (\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Sigma}_k^{-1})^{-1}$ y $m_k = V_k(\boldsymbol{\Lambda}^{-1}\mu + \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k)$, entonces, $p(\boldsymbol{\psi}_k \mid resto) \sim N(m_k, V_k)$

---

Desarrollo de $p(\sigma_k^2)$

$$p(\sigma_k^2) \propto (\sigma_k^2)^{-\alpha-1}\text{exp}\Big[-\frac{\gamma}{\sigma_k^2}\Big]$$

**A nivel de aula**:

$$\begin{aligned}
p(\sigma_k^2 \mid resto) &\propto p(\sigma_k^2)\prod_{j=1}^{n_k}\prod_{i=1}^{n_{jk}} p(y_{ijk} \mid resto) \\
&\propto \Big((\sigma_k^2)^{-\alpha-1}\text{exp}\Big[-\frac{\gamma}{\sigma_k^2}\Big]\Big)\prod_{j=1}^{n_k}\prod_{i=1}^{n_{jk}}(\sigma_k^2)^{-1/2}\text{exp}\Bigg[-\frac{1}{2}\frac{(y_{ijk}-\boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk})^2}{\sigma_k^2}\Bigg] \\
&\propto \Big((\sigma_k^2)^{-\alpha-1}\text{exp}\Big[-\frac{\gamma}{\sigma_k^2}\Big]\Big)(\sigma_k^2)^{-\frac{1}{2}\sum_{j=1}^{n_k}(n_{jk})}\prod_{j=1}^{n_k}\prod_{i=1}^{n_{jk}}\text{exp}\Bigg[-\frac{1}{2}\frac{(y_{ijk}-\boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk})^2}{\sigma_k^2}\Bigg] \\
&\propto \Big((\sigma_k^2)^{-\alpha-1}\text{exp}\Big[-\frac{\gamma}{\sigma_k^2}\Big]\Big)(\sigma_k^2)^{-\frac{1}{2}\sum_{j=1}^{n_k}(n_{jk})}\text{exp}\Bigg[\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\Bigg(-\frac{1}{2}\frac{(y_{ijk}-\boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk})^2}{\sigma_k^2}\Bigg)\Bigg],
\end{aligned}$$

Sea $N_k = \sum_{j=1}^{n_k}(n_{jk})$

$$\begin{aligned}
p(\sigma_k^2 \mid resto) &\propto (\sigma_k^2)^{-\alpha-\frac{1}{2}N_k-1}\text{exp}\Bigg[-\frac{\gamma}{\sigma_k^2}+\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\Bigg(-\frac{1}{2}\frac{(y_{ijk}-\boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk})^2}{\sigma_k^2}\Bigg)\Bigg] \\
&\propto (\sigma_k^2)^{-\alpha-\frac{1}{2}N_k-1}\text{exp}\Bigg[-\frac{1}{\sigma_k^2}\Bigg(\gamma+\frac{1}{2}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\Big((y_{ijk}-\boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk})^2\Big)\Bigg)\Bigg] \qquad (92a)
\end{aligned}$$

Por tanto,

$$p(\sigma_k^2 \mid resto) \sim IG\Big(\alpha_k, \gamma_k\Big) \qquad (93)$$

con $\alpha_k = \alpha + \frac{1}{2}N_k$ y $\gamma_k = \gamma + \frac{1}{2}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\Big((y_{ijk}-\boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{jk})^2\Big)$

**A nivel de establecimiento educativo**:

$$
\begin{aligned}
p(\sigma_k^2 \mid resto) &\propto p(\sigma_k^2) \prod_{j=1}^{n_k} \prod_{i=1}^{n_{jk}} p(y_{ijk} \mid resto) \\
&\propto \Big( (\sigma_k^2)^{-\alpha-1} \exp\Big[ -\frac{\gamma}{\sigma_k^2} \Big] \Big) \prod_{j=1}^{n_k} \prod_{i=1}^{n_{jk}} (\sigma_k^2)^{-1/2} \exp\left[ -\frac{1}{2} \frac{(y_{ijk} - \boldsymbol{X}_{ijk}^T \boldsymbol{\beta}_k)^2}{\sigma_k^2} \right] \\
&\propto \Big( (\sigma_k^2)^{-\alpha-1} \exp\Big[ -\frac{\gamma}{\sigma_k^2} \Big] \Big) (\sigma_k^2)^{-\frac{1}{2}\sum_{j=1}^{n_k}(n_{jk})} \prod_{j=1}^{n_k} \prod_{i=1}^{n_{jk}} \exp\left[ -\frac{1}{2} \frac{(y_{ijk} - \boldsymbol{X}_{ijk}^T \boldsymbol{\beta}_k)^2}{\sigma_k^2} \right] \\
&\propto \Big( (\sigma_k^2)^{-\alpha-1} \exp\Big[ -\frac{\gamma}{\sigma_k^2} \Big] \Big) (\sigma_k^2)^{-\frac{1}{2}\sum_{j=1}^{n_k}(n_{jk})} \exp\left[ \sum_{j=1}^{n_k} \sum_{i=1}^{n_{jk}} \left( -\frac{1}{2} \frac{(y_{ijk} - \boldsymbol{X}_{ijk}^T \boldsymbol{\beta}_k)^2}{\sigma_k^2} \right) \right],
\end{aligned}
$$

Sea $N_k = \sum_{j=1}^{n_k}(n_{jk})$

$$
\begin{aligned}
p(\sigma_k^2 \mid resto) &\propto (\sigma_k^2)^{-\alpha-\frac{1}{2}N_k-1} \exp\left[ -\frac{\gamma}{\sigma_k^2} + \sum_{j=1}^{n_k} \sum_{i=1}^{n_{jk}} \left( -\frac{1}{2} \frac{(y_{ijk} - \boldsymbol{X}_{ijk}^T \boldsymbol{\beta}_k)^2}{\sigma_k^2} \right) \right] \\
&\propto (\sigma_k^2)^{-\alpha-\frac{1}{2}N_k-1} \exp\left[ -\frac{1}{\sigma_k^2} \left( \gamma + \frac{1}{2} \sum_{j=1}^{n_k} \sum_{i=1}^{n_{jk}} \left( (y_{ijk} - \boldsymbol{X}_{ijk}^T \boldsymbol{\beta}_k)^2 \right) \right) \right] \qquad (95a)
\end{aligned}
$$

Por tanto,

$$
p(\sigma_k^2 \mid resto) \sim IG\Big( \alpha_k, \gamma_k \Big) \qquad (96)
$$

con $\alpha_k = \alpha + \frac{1}{2}N_k$ y $\gamma_k = \gamma + \frac{1}{2}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}} \left( (y_{ijk} - \boldsymbol{X}_{ijk}^T \boldsymbol{\beta}_k)^2 \right)$

---

Las dos restantes tienen un desarrollo exactamente igual al del modelo jerárquico de tres niveles, tanto para nivel aula como para nivel establecimiento educativo:

$$
\begin{aligned}
p(\alpha \mid resto) &= p(\alpha) p(\sigma_k^2 \mid resto) \\
&\propto \exp(-\alpha a_0) \prod_{k=1}^{M} IGamma(\sigma_k^2 \mid \alpha, \gamma) \\
&= \exp(-\alpha a_0) \prod_{k=1}^{M} \left( \frac{\gamma^\alpha}{\Gamma(\alpha)} (\sigma_k^2)^{-\alpha-1} \exp\Big[ -\frac{\gamma}{\sigma_k^2} \Big] \right) \\
&\propto \exp(-\alpha a_0) \left[ \frac{\gamma^{M\alpha}}{\Gamma(\alpha)^M} \prod_{k=1}^{M} (\sigma_k^2)^{-(\alpha+1)} \right] \qquad (97a)
\end{aligned}
$$

$$
\begin{aligned}
p(\gamma \mid resto) &\propto \prod_{k=1}^{M} IGamma(\sigma_k^2 \mid \alpha, \gamma) Gamma(\gamma \mid b_0, c_0) \\
&= \prod_{k=1}^{M} \left( \frac{\gamma^{\alpha}}{\Gamma(\alpha)} (\sigma_k^2)^{-\alpha-1} \exp\Big[ -\frac{\gamma}{\sigma_k^2} \Big] \right) \left[ \frac{c_0^{b_0}}{\Gamma(b_0)} \gamma^{(b_0-1)} \exp(-c_0\gamma) \right] \\
&\propto \gamma^{M\alpha} \exp\Big[ -\gamma \sum_{k=1}^{M} \frac{1}{\sigma_k^2} \Big] \gamma^{(b_0-1)} \exp(-c_0\gamma) \\
&= \gamma^{b_0+M\alpha-1} \exp\Big[ -\Big( c_0 + \sum_{k=1}^{M} \frac{1}{\sigma_k^2} \Big) \gamma \Big] \qquad (98a)
\end{aligned}
$$

## Anexo C. Inferencia de las distribuciones condicionales conjuntas del modelo base

Se desarrolla la verosimilitud de la ecuación (1) para una institución educativa cualquiera k:

$$p(y_{ijk} \mid resto) = \prod_{j=1}^{n_k}\left[\prod_{i=1}^{n_{jk}} \frac{1}{\sqrt{2\pi\sigma^2}}\exp -\frac{1}{2}\Big(\frac{y_{ijk}-\theta_k}{\sigma}\Big)^2\right] \tag{99a}$$

$$= \prod_{j=1}^{n_k}\left[(2\pi\sigma^2)^{-(n_{jk}/2)}\exp\Big[-\frac{1}{2\sigma^2}\sum_{i=1}^{n_{jk}}\Big(y_{ijk}-\theta_k\Big)^2)\Big]\right] \tag{99b}$$

$$= (2\pi\sigma^2)^{-(\sum_{j=1}^{n_k}(n_{jk})/2)}\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n_k}\Big(\sum_{i=1}^{n_{jk}}(y_{ijk}-\theta_k)^2\Big)\right] \tag{99c}$$

$$= (2\pi\sigma^2)^{-(\sum_{j=1}^{n_k}(n_{jk})/2)}\exp\left[-\frac{1}{2\sigma^2}\Big(\sum_{j=1}^{n_k}\Big(\sum_{i=1}^{n_{jk}}y_{ijk}^2 - 2\theta\sum_{i=1}^{n_{jk}}y_{ijk} + n_{jk}\theta_k^2\Big)\Big)\right] \tag{99d}$$

$$= (2\pi\sigma^2)^{-(\sum_{j=1}^{n_k}(n_{jk})/2)}\exp\left[-\frac{1}{2\sigma^2}\Big(\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}y_{ijk}^2 - 2\theta_k\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}y_{ijk} + \sum_{j=1}^{n_k}n_{jk}\theta_k^2\Big)\right] \tag{99e}$$

La distribución posterior, es decir, la actualización de las creencias acerca de los parámetros una vez se hallan observado los datos es $p(\theta_k, \sigma^2 \mid y_{ijk})$.

Por la definición dada en la ecuación (7) y la independencia de los dos parámetros:

$$p(\theta_k \mid \sigma^2) = p(\theta_k)$$

$$= \frac{1}{\sqrt{2\pi\tau_0^2}}\exp\left[-\frac{1}{2\tau_0^2}(\theta_k-\mu_0)^2\right] \tag{100a}$$

$$\propto \exp\Big[-\frac{1}{2\tau_0^2}(\theta_k-\mu_0)^2\Big], \tag{100b}$$

Se está en disposición de desarrollar la distribución posterior para cada $\theta_k$:

$$p(\theta_k \mid resto) \propto \exp\left[-\frac{1}{2\sigma^2}\Big(\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}y_{ijk}^2 - 2\theta_k\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}y_{ijk} + \sum_{j=1}^{n_k}n_{jk}\theta_k^2\Big)\right]\exp\Big[-\frac{1}{2\tau_0^2}(\theta_k-\mu_0)^2\Big]$$

$$= \exp\left[-\frac{1}{2\sigma^2}\Big(\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}y_{ijk}^2 - 2\theta_k\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}y_{ijk} + \sum_{j=1}^{n_k}n_{jk}\theta_k^2\Big) - \frac{1}{2\tau_0^2}\Big(\theta_k^2 - 2\theta_k\mu_0 + \mu_0^2\Big)\right] \tag{101a}$$

Ordenando la expresión del exponente e ignorando el -1/2 por el momento:

$$\theta_k^2\left(\frac{1}{\tau_0^2}+\frac{\sum_{j=1}^{n_k} n_{jk}}{\sigma^2}\right)-2\theta_k\left(\frac{\mu_0}{\tau_0^2}+\frac{\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}} y_{ijk}}{\sigma^2}\right)+\left(\frac{\mu_0^2}{\tau_0^2}+\frac{\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}} y_{ijk}^2}{\sigma^2}\right) \tag{102}$$

Si cada multiplicando es rebautizado del siguiente modo:

$$a=\left(\frac{1}{\tau_0^2}+\frac{\sum_{j=1}^{n_k} n_{jk}}{\sigma^2}\right),$$

$$b=\left(\frac{\mu_0}{\tau_0^2}+\frac{\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}} y_{ijk}}{\sigma^2}\right),$$

y

$$c=\left(\frac{\mu_0^2}{\tau_0^2}+\frac{\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}} y_{ijk}^2}{\sigma^2}\right),$$

entonces, la ecuación (101a) es $\exp\big[-\frac{1}{2}(a\theta_k^2-2b\theta_k+c)\big]$.

Sean entonces las siguientes definiciones:

$$B_k^2=\frac{1}{a}=\frac{1}{\frac{1}{\tau_0^2}+\frac{\sum_{j=1}^{n_k} n_{jk}}{\sigma^2}}=\Big[\frac{\tau_0^2\sigma^2}{\sigma^2+\tau_0^2\sum_{j=1}^{n_k} n_{jk}}\Big] \tag{103a}$$

$$A_k=\frac{b}{a}=\frac{\frac{1}{\tau_0^2}\mu_0+\frac{\sum_{j=1}^{n_k} n_{jk}}{\sigma^2}\bar{y}_k}{\frac{1}{\tau_0^2}+\frac{n_{jk}}{\sigma^2}}=\Big[\frac{1}{\tau_0^2}\mu_0+\frac{\sum_{j=1}^{n_k} n_{jk}}{\sigma_0^2}\bar{y}_k\Big]\Big[\frac{\tau_0^2\sigma^2}{\tau_0^2\sigma^2+\sum_{j=1}^{n_k} n_{jk}}\Big], \tag{103b}$$

Corresponden a la varianza y la media, respectivamente, de la distribución normal posterior del parámetro $\theta_k$.

---

Condicionado a que $\sigma^2$ es igual para todas las instituciones educativas:

$$p(\sigma^2 \mid resto)\propto\prod_{k=1}^{M}\left(\prod_{j=1}^{n_k}\left(\prod_{i=1}^{n_{jk}}\Big(p(y_{ijk}\mid\theta_k,\sigma^2)\Big)\right)\right)p(\sigma^2) \tag{104}$$

Por la definición de la ecuación (3)

$$p(\sigma^2)=\frac{z_0^{y_0}}{\Gamma\Big(y_0\Big)}(\sigma^2)^{-y_0-1}\exp\left[-\frac{z_0}{\sigma^2}\right]$$

$$\propto(\sigma^2)^{-y_0-1}\exp\left[-\frac{z_0}{\sigma^2}\right] \tag{105a}$$

Por su parte, aplicando 1 y 3:

$$
\begin{aligned}
p(\sigma^2 \mid resto) &\propto \prod_{k=1}^{M}\left(\prod_{j=1}^{n_k}\left(\prod_{i=1}^{n_{jk}}\frac{1}{\sqrt{2\pi\sigma^2}}\left[\exp\left(-\frac{1}{2\sigma^2}\Big(y_{ijk}-\theta_k\Big)^2\right)\right]\right)\right) * (\sigma^2)^{-y_0-1}\exp\left[-\frac{z_0}{\sigma^2}\right] \\
&= \prod_{k=1}^{M}\left(\prod_{j=1}^{n_k}\left((2\pi\sigma^2)^{-(n_{jk}/2)}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n_{jk}}\Big(y_{ijk}-\theta_k\Big)^2\right]\right)(\sigma^2)^{-y_0-1}\exp\left[-\frac{z_0}{\sigma^2}\right]\right) \\
&\propto \prod_{k=1}^{M}\left((\sigma^2)^{-\frac{1}{2}(\sum_{j=1}^{n_k} n_{jk})}\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\Big(y_{ijk}-\theta_k\Big)^2\right](\sigma^2)^{-y_0-1}\exp\left[-\frac{z_0}{\sigma^2}\right]\right) \\
&\propto (\sigma^2)^{-\frac{1}{2}\left(\sum_{k=1}^{M}(\sum_{j=1}^{n_k} n_{jk})\right)}\exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{M}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\Big(y_{ijk}-\theta_k\Big)^2\right](\sigma^2)^{-y_0-1}\exp\left[-\frac{z_0}{\sigma^2}\right] \\
&\propto (\sigma^2)^{-\left(y_0+\frac{1}{2}\left(\sum_{k=1}^{M}(\sum_{j=1}^{n_k} n_{jk})\right)+1\right)}\exp\left(-\frac{1}{\sigma^2}\left[z_0+\frac{1}{2}\sum_{k=1}^{M}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\Big(y_{ijk}-\theta_k\Big)^2\right]\right)
\end{aligned}
\tag{106a}
$$

que tiene la forma de una distribución Gamma inversa, siendo

$$
C_k = y_0 + \frac{1}{2}\sum_{k=1}^{M}\sum_{j=1}^{n_k} n_{jk} \tag{107}
$$

y

$$
D_k = z_0 + \frac{1}{2}\sum_{k=1}^{M}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}\left(y_{ijk}-\theta_k\right)^2 \tag{108}
$$

Implica que la ecuación (106a) se escriba de manera sucinta así:

$$
p(\sigma^2 \mid \theta_k, y_{ijk}, y_0, z_0) \propto (\sigma^2)^{-C_k-1}\exp\left[-\frac{1}{\sigma^2}D_k\right] \tag{109}
$$

## Anexo D. Algoritmo Metrópolis

En el modelo jerárquico de tres niveles, tanto $\alpha$ como $\lambda$ tienen distribuciones posteriores de las que no tenemos funciones preprogramadas de las que muestrear. Recordemos que en las distribuciones compuestas la distribución marginal de alguno de los parámetros se halla integrando respecto a los restantes parámetros, no obstante, en este tipo de distribuciones no cerradas es difícil calcular tales integrales. Pero es posible generar una gran colección de valores $\alpha^{(1)}, \alpha^{(2)}, \cdots, \alpha^{(S)} \sim p(\alpha \mid y)$, para obtener una aproximación Montecarlo a los valores de la respectiva distribución.

Metrópolis es una solución algorítmica.

Sea mostrado para $\alpha$ del modelo jerárquico de tres niveles (ecuación 65a).

$$p(\alpha \mid resto) \propto \frac{\eta^{(M\alpha)} e^{(-\alpha a_0)}}{\Gamma(\alpha)^M} \prod_{k=1}^{M} (\sigma_k^2)^{-(\alpha+1)} \tag{110}$$

Hoff explica que la idea es explorar el espacio muestral mediante una caminata aleatoria de, por ejemplo S pasos, tal vez 10,000, en la que se propone un valor $\alpha^*$ cercano a un valor $\alpha^{(s)}$, donde $s \epsilon \{S\}$. Se debe establecer si se acepta el nuevo valor propuesto.

Sea la distribución del valor propuesto $J(\alpha^* \mid \alpha^{(s)}) \sim N(\alpha^{(s)}, \delta^2)$. Es decir, cada $\alpha^*$ depende del valor $\alpha^{(s)}$ más una pequeña perturbación. Luego, se acepta dicha propuesta con una cierta probabilidad, en caso contrario se permanece con el valor $\alpha^{(s)}$. De cualquier manera $\alpha^{(s+1)}$ depende del valor $\alpha^{(s)}$. Es un proceso de Markov.

El valor del parámetro $\delta$ se escoge de tal modo que el tamaño de los pasos de la exploración del espacio muestral no sean ni tan pequeños que demore mucho en alcanzar una convergencia, ni tan largos que frecuentemente $\alpha^*$ se ubique lejos de la distribución posterior buscada, y por tanto sea rechazado, quedando estancado el algoritmo por largos periodos, buscando valores que no se rechacen.

La forma de aceptar o rechazar el valor $\alpha^*$ es calculando la razón $r = p(\alpha^* \mid y)/p(\alpha^{(s)} \mid y)$, la cual se conoce como la tasa de aceptación. Sea descrito el algoritmo:

1. Muestree $\alpha^*$ de $J(\alpha^* \mid \alpha^{(s)})$;
2. Calcule la tasa de aceptación r;
3. Sea $u \sim uniforme(0, 1)$;
4. Sea $\alpha^{(s+1)} = \alpha^*$ si $u \geq r$, sea $\alpha^{(s+1)} = \alpha^{(s)}$ de lo contrario.

La forma de $p(\alpha^* \mid y)$ y de $p(\alpha^{(s)} \mid y)$ esta dada por la ecuación (110). Hoff recomienda calcular log(r) en vez de r directamente para lograr una estabilidad en los cálculos.

Por tanto la ecuación (110) se transforma en:

$$\log(p(\alpha \mid resto)) \propto -(\alpha+1) \sum_{k=1}^{M} \log(\sigma_k^2) + M\alpha \log(\eta) - M\Gamma(\alpha) - \alpha a_0 \tag{111}$$

La forma de determinar si se está logrando una buena búsqueda tiene que ver con el cálculo de las autocorrelaciones entre cada par de valores. Se busca minimizar dichas autocorrelaciones, tal vez mediante simulaciones de varios posibles valores de $\delta$.

Como ya se mencionó, Hoff [11] presenta que el parámetro de forma de una distribución gamma inversa, restringiéndolo a ser un número entero, se puede modelar con una distribución previa exponencial. Esto implica que es válido muestrear de la ecuación (110) especificando solamente valores alfa enteros del 1 a n, con n = 50, ó n = 1000, según convenga.

# Anexo E. La distribución gamma inversa

## E.1 Introducción

Las funciones de distribución Gamma y Gamma inversa tienen interés práctico.

La distribución Gamma surge naturalmente en procesos para los cuales los tiempos de espera entre eventos son relevantes en una gran variedad de disciplinas que incluyen modelamientos de colas, climatología y servicios financieros. Ejemplos de eventos que pueden ser modelados por la distribución Gamma incluyen:

- La cantidad de lluvia acumulada en un embalse.
- El tamaño de los impagos de préstamos o reclamos de seguros agregados.
- El flujo de artículos a través de procesos de fabricación y distribución.
- La carga en servidores web.
- Las muchas y variadas formas de intercambio de telecomunicaciones.

La función de densidad *gamma inversa* surge con frecuencia en el análisis Bayesiano de datos normales, como previa conjugada de un parámetro de varianza desconocida., si bien también surge en comunicaciones. En los canales inalámbricos, las fluctuaciones aleatorias que afectan a las señales de radio se han dividido clásicamente en dos tipos: desvanecimiento rápido (fast fading), como resultado de la propagación por trayectos múltiples, y el sombreado (shadowing), que es causado por la presencia de objetos grandes como árboles o edificios. Con el objetivo de estudiar y mejorar el rendimiento de los sistemas de comunicación inalámbrica, se han dedicado esfuerzos considerables a la caracterización de estos dos efectos que, en muchos casos, se analizan por separado. A pesar de caracterizarlos como fenómenos diferentes, en la práctica ocurren simultáneamente, aunque a diferentes escalas de tiempo. Por lo tanto, los modelos de desvanecimiento compuesto surgieron para caracterizar el impacto combinado de estos dos efectos. En los últimos años, la distribución Gamma Inversa ha comenzado a usarse para caracterizar el sombreado, motivados por el hecho de que admite una formulación matemática relativamente simple.[5]

En actuaría o análisis de riesgos, la distribución Gamma es usada en el modelado de tamaños de reclamo pequeño y moderado en compañías de seguros, y la distribución Gamma inversa se aplica en el modelado de tamaño de reclamos grandes[6] como consecuencia de tener una cola más pesada que la distribución Gamma. Una mezcla de Gamma y Gamma inversa puede ser usada como modelo estadístico aplicable para reclamos de todos los tamaños.

## E.2 Transformada inversa

Para obtener la densidad de una variable aleatoria continua transformada, se debe tener en cuenta la tasa de cambio de la variable aleatoria original con respecto a la nueva variable aleatoria. Sea Y = 1/X. Entonces, también X = 1/Y, por ende el valor absoluto del jacobiano de la transformación[7] es:

$$\left|\frac{dx}{dy}(y)\right| = \left|\frac{d}{dy}\left(\frac{1}{y}\right)\right| = \left|-\frac{1}{y^2}\right| = \frac{1}{y^2}$$

[5] Composite Fading Models based on Inverse Gamma Shadowing: Theory and Validation. Pablo Ramírez-Espinosa and F. Javier López-Martínez. https://arxiv.org/pdf/1905.00069.pdf v3 2020

[6] Modelling Insurance Claim Sizes using the Mixture of Gamma & Reciprocal Gamma Distributions. Ying Ni. 2015

[7] https://math.stackexchange.com/questions/856654/why-absolute-values-of-jacobians-in-change-of-variables-for-multiple-integrals-b

Sea $X \sim Gamma(\alpha, \beta)$, entonces

$$\begin{aligned} f(Y;\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{y}\right)^{\alpha-1} exp\Big[\frac{-\beta}{y}\Big]\frac{1}{y^2} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{y}\right)^{\alpha-1} exp\Big[\frac{-\beta}{y}\Big]\left(\frac{1}{y}\right)^{2} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{y}\right)^{\alpha+1} exp\Big[\frac{-\beta}{y}\Big] \end{aligned}$$

O, escribiéndolo de manera más simple, se dice que una variable y tiene una función de distribución Gamma Inversa si:

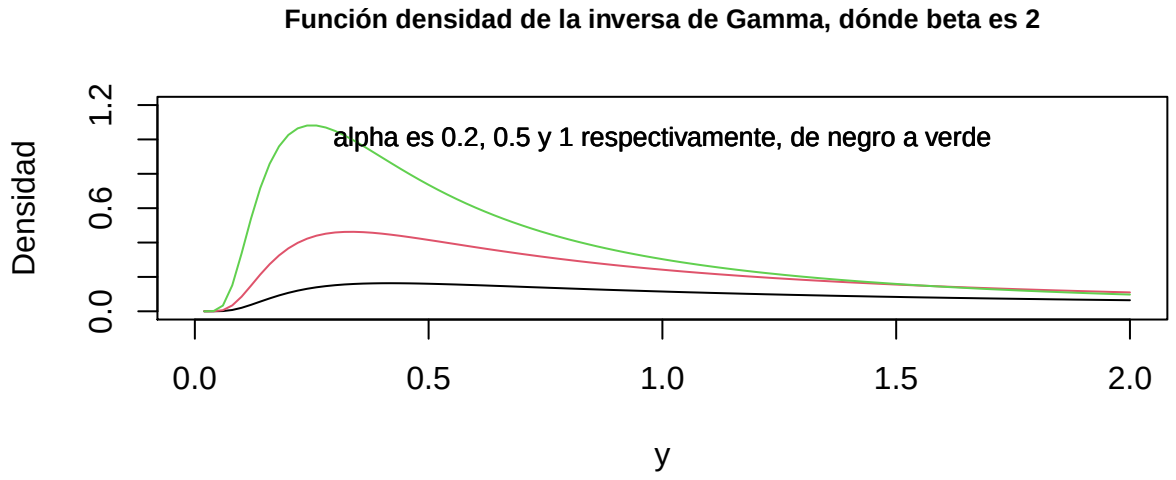
$$f(Y;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-\alpha-1} exp\Big[-\frac{\beta}{y}\Big]$$

Si se parametriza el segundo parámetro como de razón $\lambda = 1/\beta$:

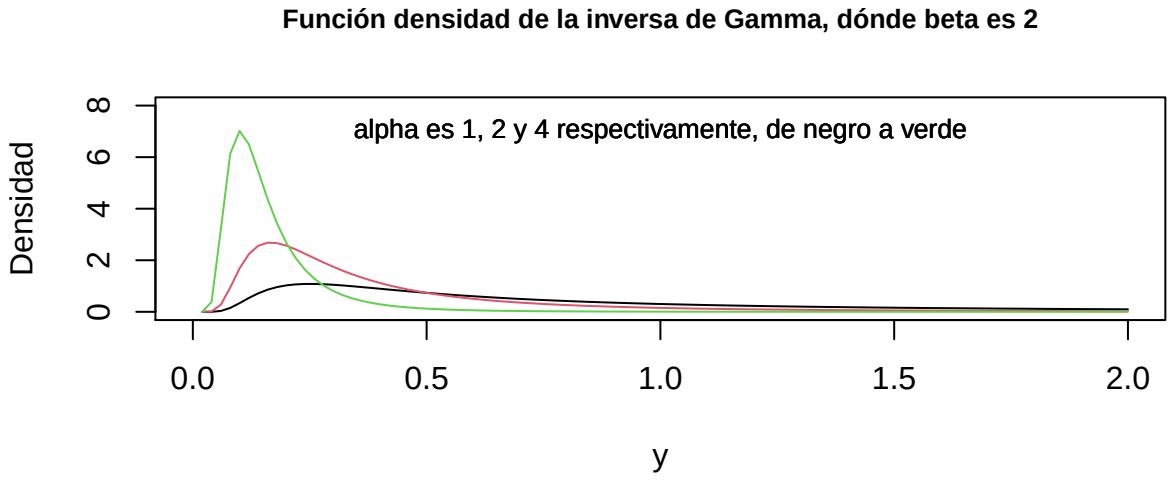
$$f(Y;\alpha,\lambda) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{-\alpha-1} exp\Big[-\frac{1}{y\beta}\Big]$$

Asúmase inicialmente la forma $f(Y;\alpha,\beta)$.

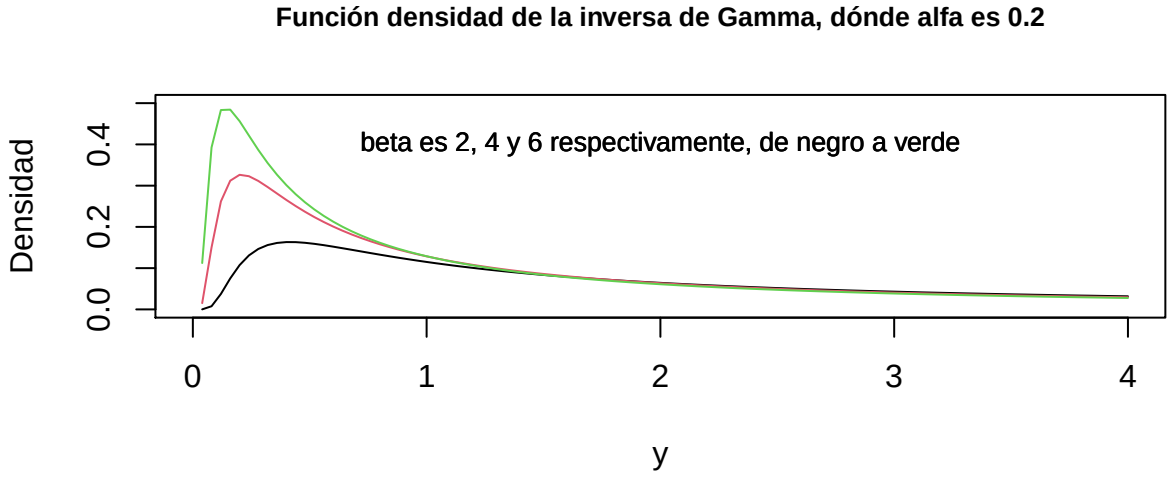
$\alpha$ controla la altura de la curva. A mayor $\alpha$, mayor altura de la curva. Es el parámetro de forma.



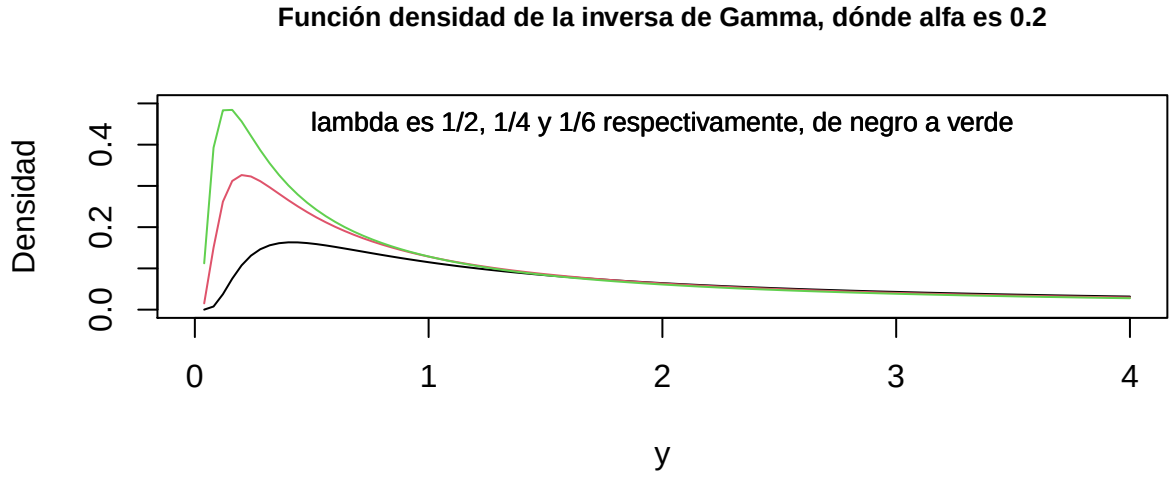
Si $\alpha \geq 1$:




$\beta$ es el parámetro de escala y controla la precisión. Mayor escala, mayor precisión.



$\lambda = 1/\beta$ se denomina el parámetro de rata o razón y controla la dispersión. Mayor rata, mayor dispersión.




Es equivalente a un $\beta$ de 2, 4 y 6 respectivamente, de negro a verde, es decir, igual a la gráfica anterior.

Se observa que la inversa de Gamma tiene colas más pesadas que Gamma.

### E.3 Momentos de la Inversa de Gamma

En general:

$$\begin{aligned} E(Y^n) &= \int_0^\infty y^n \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-\alpha-1} \exp(-\beta/y) dy \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty y^n y^{-\alpha-1} \exp(-\beta/y) dy \end{aligned} \tag{113a}$$

Antes de continuar, obsérvese que, si iniciamos con $z$ y luego transformamos $z = \beta/y$, se puede aplicar el mismo razonamiento con que inició el presente literal. El valor absoluto del jacobiano de la transformación es $dz = \beta/y^2 dy$, y la siguiente igualdad se cumple:

$$\begin{aligned} \beta^{n-\alpha}\Gamma(\alpha - n) &= \beta^{n-\alpha} \int_0^\infty z^{-(n-\alpha)-1} exp[-z] dz \\ &= \beta^{n-\alpha} \int_0^\infty \frac{\beta^{-(n-\alpha)-1}}{y^{-(n-\alpha)-1}} \frac{\beta}{y^2} exp[-\beta/y] dy \\ &= \frac{\beta^{(n-\alpha)}}{\beta^{(n-\alpha)}} \int_0^\infty \frac{1}{y^{-n+\alpha+1}} exp[-\beta/y] dy \\ &= \int_0^\infty y^n y^{-\alpha-1} exp[-\beta/y] dy \end{aligned}$$

Por ende, la ecuación (113a) se reescribe como:

$$\begin{aligned} E(Y^n) &= \frac{\beta^\alpha \Gamma(\alpha - n)}{\Gamma(\alpha)(\beta^{\alpha-n})} \\ &= \frac{\beta^n \Gamma(n - \alpha)}{\Gamma(\alpha)} \\ &= \frac{\beta^n}{(\alpha-1)(\alpha-2)\dots(\alpha-n)} \end{aligned}$$

con $\alpha > n$ para que sea positivo.

Por tanto,

$E(Y) = \frac{\beta}{(\alpha-1)}$ con $\alpha > 1$.

$$\begin{aligned} Var(Y) &= \frac{\beta^2}{(\alpha-1)(\alpha-2)} - \left(\frac{\beta}{(\alpha-1)}\right)^2 \\ &= \frac{\beta^2}{(\alpha-1)^2(\alpha-2)},\ si\ \alpha > 2. \end{aligned}$$

Sobre un parámetro de razón:

$E(Y) = \frac{\lambda}{(\alpha-1)}$ y $var(Y) = \frac{\lambda^2}{(\alpha-1)^2(\alpha-2)}$

Por tanto, parametrizando con $\lambda$, a mayor $\alpha$ y $\beta$, mayor precisión.

# Anexo F. Validación de los modelos para los índices de Habilidad y Respeto

## F.1 Habilidades comunicativas

### F.1.1 Modelo base

El cálculo de la Log-verosimilitud del MCMC para el modelo base es:

$$logP\Big(\tilde{y} \mid \theta^{(s)}, (\sigma^2)^{(s)}\Big) = \Sigma_{k=1}^{M}\Sigma_{j=1}^{n_k}\Sigma_{i=1}^{n_{jk}} N(\theta_k, \sigma^2)$$

Dónde el superíndice (s) representa la iteración que corresponde.

$$s \in \{1, 2, 3, \ldots, S\}$$

S es el número total de iteraciones.

La siguiente gráfica presenta la convergencia del modelamiento para $logP\Big(\tilde{y} \mid \theta^{(s)}, (\sigma^2)^{(s)}\Big)$.

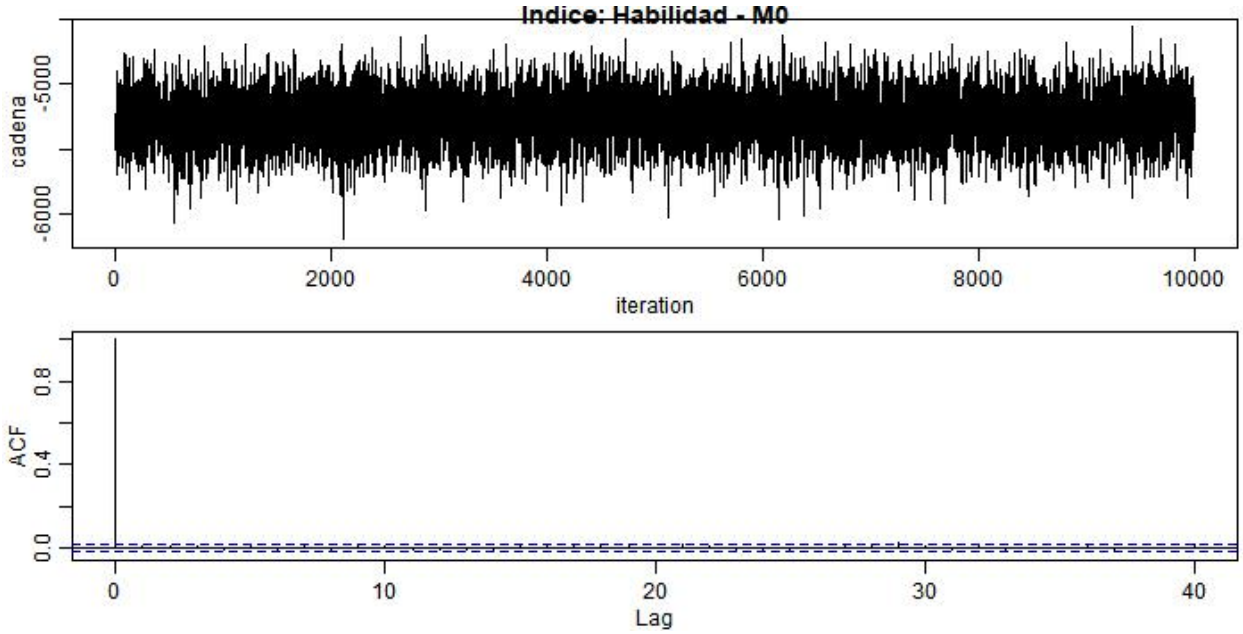



Figura 47: Convergencia del modelamiento para logP

Se realizaron 52,000 iteraciones, con un calentamiento de la serie de 2,000 iteraciones y guardando los resultados cada 5 iteraciones con el objeto de obtener finalmente 10,000 muestras. Se observa un muy buen comportamiento, con nula auto-regresión entre las muestras. Bajo pruebas formales, realizadas con el paquete *coda* de R, la cadena converge.

El tamaño efectivo de $\theta$ y $\sigma^2$ es 10,050 y 9,999 respectivamente.

Cuadro 15: Errores estándar del modelo base

| Distribución | Error estándar |
|---|---|
| Theta | 0.0004094 |
| Sigma2 | 0.0000406 |

Los errores estándar son adecuados.

La validación cruzada realizada sobre el modelo base obtuvo un error cuadrático medio (AMSE) de 0.988.

El DIC del modelos es -756,356.21.

Para evaluar la consistencia interna se determina el *predictive posterior p-value (ppp)*. Es de esperar que dicho valor se sitúe entre el 2.5 % y el 97.5 % para que sea consistente.

Cuadro 16: Porcentaje de establecimientos educativos, en el modelo base, con un predictive posterior p-value dentro de los rangos esperados

| | Media | Mediana | Desviación estándar |
|---|---|---|---|
| Establecimiento educativo | 100 | 60.7 | 0 |

El cuadro 16 presenta el porcentaje de establecimientos educativos cuyo ppp-value está dentro de los límites especificados. Se observa que el modelo tiene problemas en representar bien la desviación estándar. Y el valor de la mediana también es bajo. Esto todavía permite responder las preguntas que se plantearon al inicio del trabajo por cuanto se refieren todas a la media.

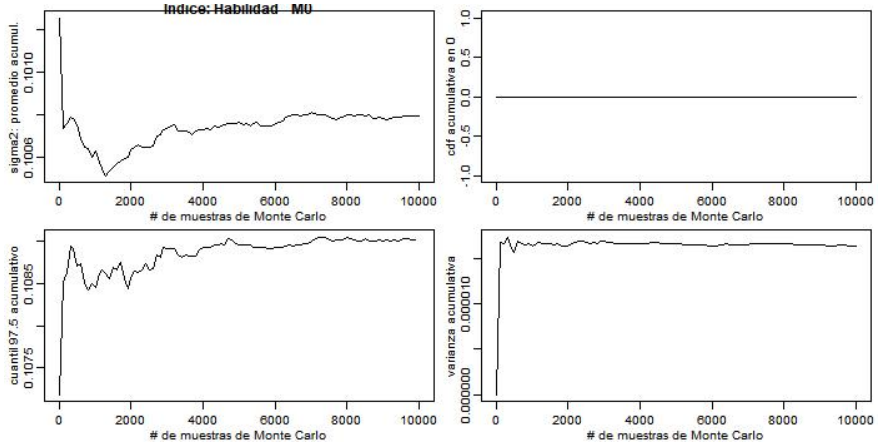
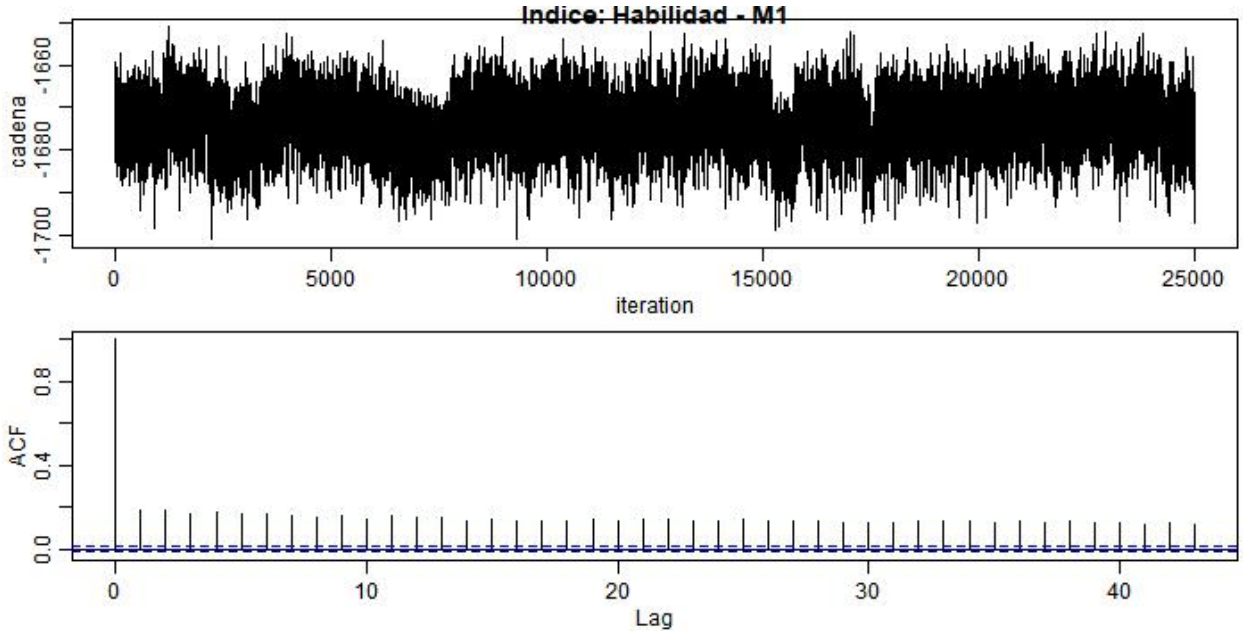



Figura 48: Convergencia del parámetro sigma2

La Figura 48 presenta que se llegó a la convergencia en el parámetro $\sigma^2$ del modelo base, por tanto, no es problema atribuible a un bajo número de iteraciones, sino a que la especificación no logra representar adecuadamente los datos.

### F.1.2 Modelo jerárquico de tres niveles




La convergencia es adecuada, pero presenta autocorrelación. Bajo pruebas formales, realizadas con el paquete *coda* de R, la cadena converge.

Obsérvese que el número de iteraciones fue de 25,000 para el modelo, como ya se explicó.

El comportamiento de las variables de interés: $\theta_{jk}$ y $\sigma_k^2$ es bueno, con muestras efectivas de 13,694 y 21,016. También los tamaños de muestra efectiva de $\mu_k$, $\gamma$ y $\kappa^2$ fueron buenos: 22,520, 24,306 y 24,999 respectivamente. Corresponde al modelamiento de la media de $\theta$. Pero los tamaños de muestra efectiva de $\tau^2$, $\lambda$ y $\xi$ fueron pésimos: 77, 59 y 39 respectivamente. Corresponden al modelamiento de la varianza de $\theta$.

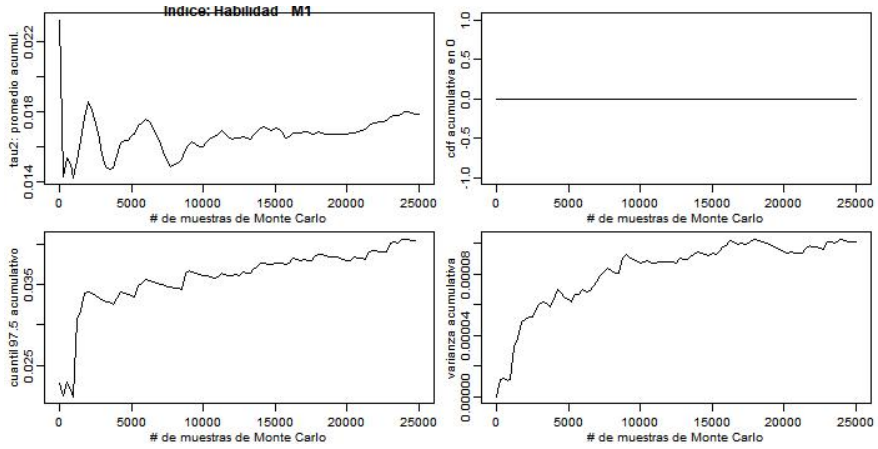



Figura 49: Convergencia del parámetro tau2

En la Figura 49 se observa cómo, a diferencia de los otros índices, el parámetro $\tau^2$ del nivel 2 logra estabilización (convergencia).

Cuadro 17: Errores estándar del modelo jerárquico de tres niveles

| Distribución | Error estándar |
|---|---|
| Theta | 0.0011321 |
| Sigma2 | 0.0008270 |
| Mu | 0.0008705 |
| Tau2 | 0.0011446 |
| Alpha | 0.2518747 |
| Eta | 0.1970242 |
| Gamma | 0.0003671 |
| Kappa2 | 0.0001260 |
| Lambda | 34.4133823 |
| Xi | 0.9864313 |

Se observa un error estándar aparentemente alto para $\alpha$, $\eta$, $\lambda$ y $\xi$. No obstante, en relación a las medias no son valores altos (la media entre el EE es, respectivamente, 96, 94.1, 15, 8.9).

El modelo jerárquico de tres niveles obtuvo un AMSE de 0.98. Una desmejora del 0.8 % respecto al modelo base.

El DIC del modelo es -31,930.58.

Cuadro 18: Porcentaje de establecimientos educativos, en el modelo jerárquico de tres niveles, con un predictive posterior p-value dentro de los rangos esperados

| | Media | Mediana | Desviación estándar |
|---|---|---|---|
| Aula | 100 | 100 | 94.2 |
| Establecimiento educativo | 100 | 100 | 96.4 |

El modelo jerárquico de tres niveles obtiene un desempeño inferior, medido mediante el DIC, al modelo base. No obstante, es consistente para las tres estadísticas tanto a nivel de establecimiento educativo, como de aula.

### F.1.1 Modelo jerárquico de dos niveles, con covariables

El cálculo de la Log-verosimilitud del MCMC para los modelos jerárquicos de dos niveles es:

**A nivel de aula**

$$logP\left(\tilde{y}_{ijk} \mid \beta_{jk}^{(s)}, (\sigma_k^2)^{(s)}\right) = \Sigma_{k=1}^{M}\Sigma_{j=1}^{n_k}\Sigma_{i=1}^{n_{jk}} N(\beta_{jk}^{(s)} X_{ijk}^{T}, (\sigma_k^2)^{(s)})$$

**A nivel de establecimiento educativo**

$$logP\left(\tilde{y}_{ijk} \mid \beta_{k}^{(s)}, (\sigma_k^2)^{(s)}\right) = \Sigma_{k=1}^{M}\Sigma_{j=1}^{n_k}\Sigma_{i=1}^{n_{jk}} N(\beta_{k}^{(s)} X_{ijk}^{T}, (\sigma_k^2)^{(s)})$$

Dónde el superíndice (s) representa la iteración que corresponde.

$$(s) \in \{1, 2, 3, \ldots, S\}$$

S es el número total de iteraciones.

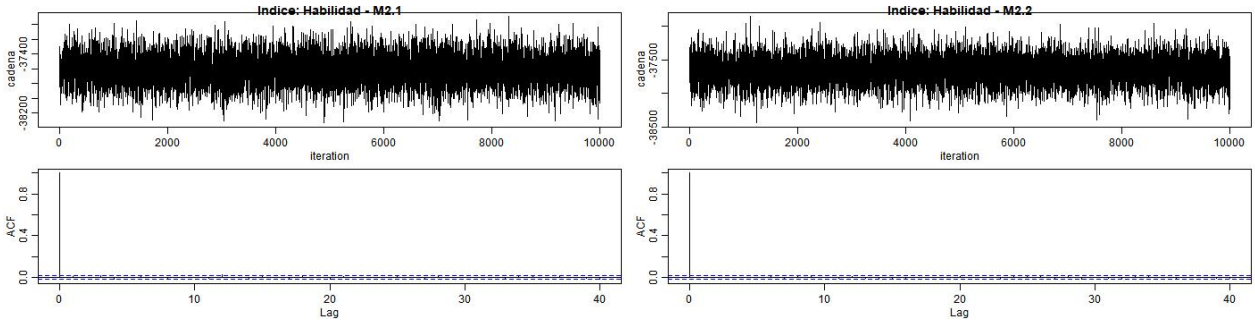



Figura 50: Convergencia de los modelos para el indice de Habilidades comunicativas

Las gráficas de la Figura 50 presentan la cadena de la logverosimilitud para cada uno de los modelos. Las dos opciones del modelo convergieron de manera adecuada. Para ambos se utilizó una configuración de iteraciones de MCMC: 102,500 iteraciones, con un calentamiento de la serie de 2,500 iteraciones, guardada de los resultados cada 10 iteraciones con el objeto de obtener 10,000 muestras.

Cuadro 19: Errores estándar del modelo M2.1 para el quinto establecimiento educativo

| Distribución | Error estándar |
|---|---|
| Beta_0 | 0.0003980 |
| Beta_1 | 0.0005206 |
| Sigma2 | 0.0000045 |

Los tamaños efectivos para, por ejemplo, $\beta_0$, $\beta_1$ y $\sigma^2$ del primer modelo jerárquico de dos niveles fueron: 10,026, 9,959 y 1,836,504 respectivamente.

Cuadro 20: Errores estándar del modelo M2.2 para el quinto establecimiento educativo

| Distribución | Error estándar |
|---|---|
| Beta_0 | 0.0212980 |
| Beta_1 | 0.0213835 |
| Beta_2 | 0.0379819 |
| Beta_3 | 0.0214965 |
| Sigma2 | 0.0000046 |

Los tamaños efectivos para el segundo modelo jerárquico de dos niveles fueron: 9,889, 9,811, 9,684, 9,705 y 10,042 respectivamente.

En comparación con el modelo por sexo, el modelo por grado obtiene un mayor error estándar en sus coeficientes $\beta$ (la media de los betas entre sus respectivos EE es 31.3, 48.5, 0.1, 34.9).

El primer modelo jerárquico de dos niveles obtuvo un AMSE de 1.038 en la validación cruzada. Una mejora del 5.1 % respecto al Modelo base y del 6 % respecto al modelo jerárquico de tres niveles.

El DIC de los modelos es:

Cuadro 21: Deviance Information Criterion por modelo

| Modelo | DIC |
|---|---|
| Sexo | 53,449 |
| Grado | 62,491 |

Los modelos jerárquicos de dos niveles obtienen un DIC muy superior en valor, muy inferior en desempeño al modelo base.

A pesar del DIC mayor, los modelos jerárquicos de dos niveles permiten realizar inferencias sobre sexo y grado respectivamente, lo cual es una ganancia frente al aparente mejor desempeño de los otros dos modelos.

Cuadro 22: Porcentaje de aulas, en los modelos jerárquicos de dos niveles, con un predictive posterior p-value dentro de los rangos esperados

| | Media | Mediana | Desviación estándar |
|---|---|---|---|
| Sexo a nivel de Aula | 100 | 47.8 | 0 |
| Grado a nivel de establecimiento educativo | 100 | 35.7 | 0 |

Los modelos jerárquicos de dos niveles tienen un desempeño regular en la representación de lo que ocurre en las aulas o establecimientos educativos para las estadísticas de mediana y desviación estándar, sobre todo a nivel de desviación estándar. Pero para la media la representación es adecuada.

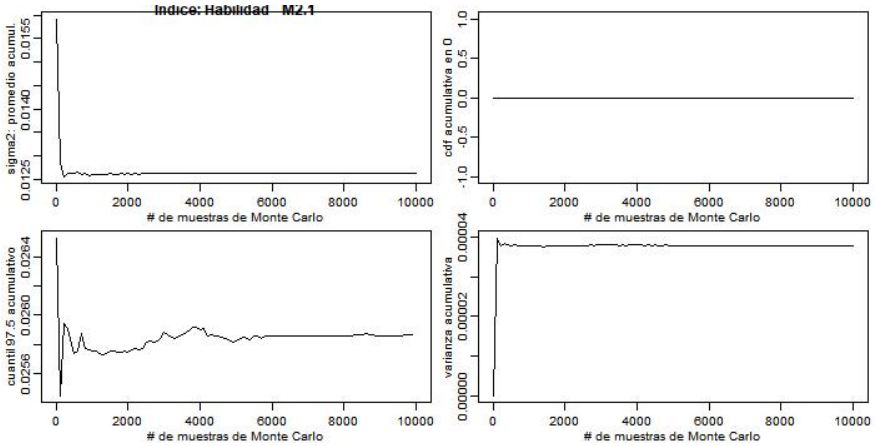



Figura 51: Convergencia del parámetro sigma2 para el primer modelo jerárquico de dos niveles

Las figuras 51 y 52 presentan que también los modelos jerárquicos de dos niveles logran convergencia del parámetro $\sigma^2$.

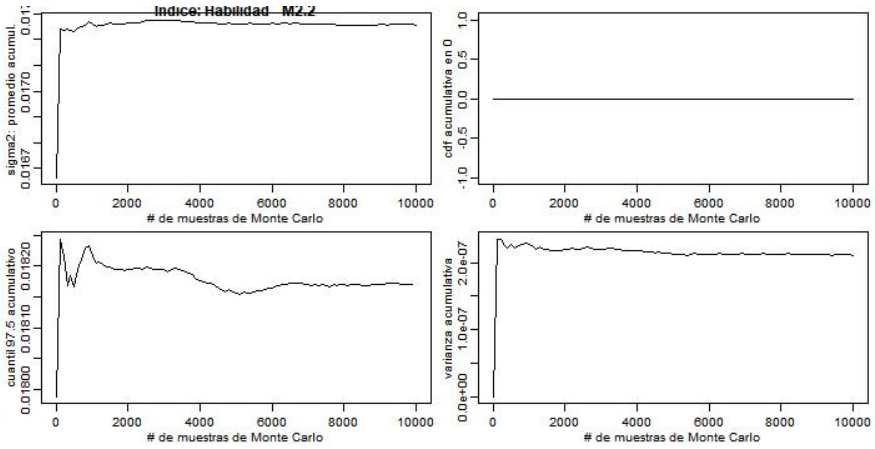



Figura 52: Convergencia del parámetro sigma2 para el segundo modelo jerárquico de dos niveles

## F.2 Respeto por los demás y por sí mismo

### F.2.1 Modelo base

El cálculo de la Log-verosimilitud del MCMC para el modelo base es:

$$logP\left(\tilde{y} \mid \theta^{(s)}, (\sigma^2)^{(s)}\right) = \Sigma_{k=1}^{M}\Sigma_{j=1}^{n_k}\Sigma_{i=1}^{n_{jk}} N(\theta_k, \sigma^2)$$

Dónde el superíndice (s) representa la iteración que corresponde.

$$s \in \{1, 2, 3, \ldots, S\}$$

S es el número total de iteraciones.

La siguiente gráfica presenta la convergencia del modelamiento para $logP\left(\tilde{y} \mid \theta^{(s)}, (\sigma^2)^{(s)}\right)$.

Se realizaron 52,000 iteraciones, con un calentamiento de la serie de 2,000 iteraciones y guardando los resultados cada 5 iteraciones con el objeto de obtener finalmente 10,000 muestras. Se observa un muy buen comportamiento, con nula auto-regresión entre las muestras. Bajo pruebas formales, realizadas con el paquete *coda* de R, la cadena converge.

El tamaño efectivo de $\theta$ y $\sigma^2$ es 10,050 y 10,000 respectivamente.

Cuadro 23: Errores estándar del modelo base

| Distribución | Error estándar |
|---|---|
| Theta | 0.0005341 |
| Sigma2 | 0.0000693 |

Los errores estándar son adecuados.

La validación cruzada realizada sobre el modelo base obtuvo un error cuadrático medio (AMSE) de 0.84.

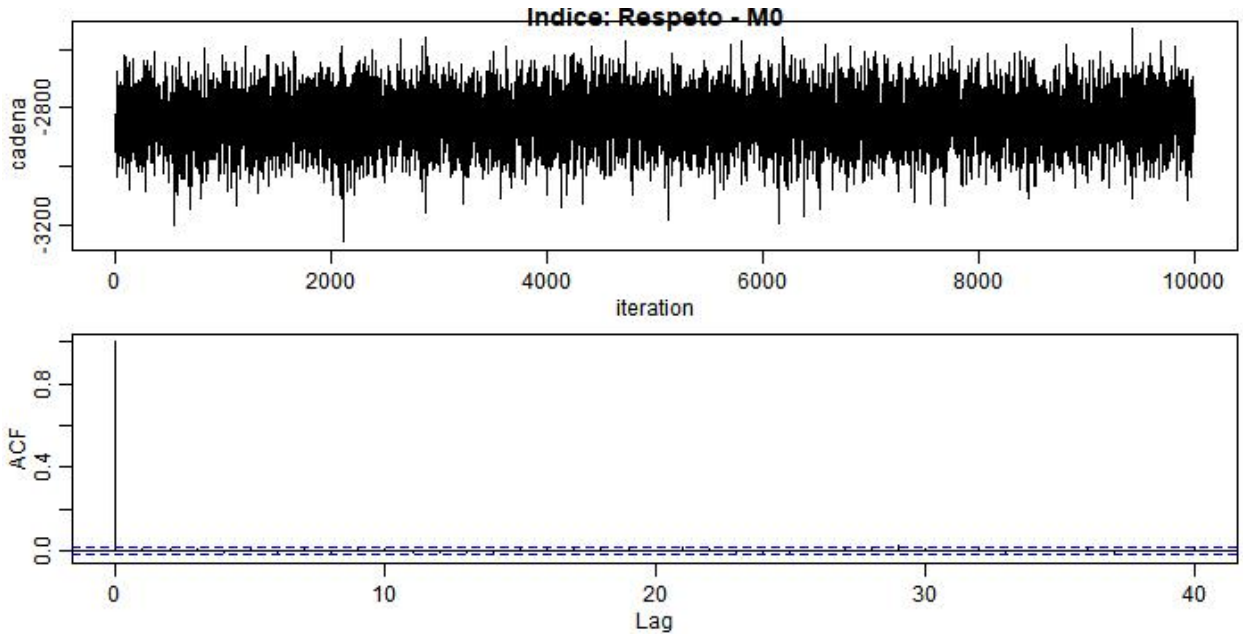



Figura 53: Convergencia del modelamiento para logP

El DIC del modelos es -756,356.21.

Para evaluar la consistencia interna se determina el *predictive posterior p-value (ppp)*. Es de esperar que dicho valor se sitúe entre el 2.5 % y el 97.5 % para que sea consistente.

Cuadro 24: Porcentaje de establecimientos educativos, en el modelo base, con un predictive posterior p-value dentro de los rangos esperados

| | Media | Mediana | Desviación estándar |
|---|---|---|---|
| Establecimiento educativo | 100 | 60.7 | 0 |

El cuadro 24 presenta el porcentaje de establecimientos educativos cuyo ppp-value está dentro de los límites especificados. Se observa que el modelo tiene problemas en representar bien la desviación estándar. Y el valor de la mediana también es bajo. Esto todavía permite responder las preguntas que se plantearon al inicio del trabajo por cuanto se refieren todas a la media.

La Figura 54 presenta que se llegó a la convergencia en el parámetro $\sigma^2$ del modelo base, por tanto, no es problema atribuible a un bajo número de iteraciones, sino a que la especificación no logra representar adecuadamente los datos.

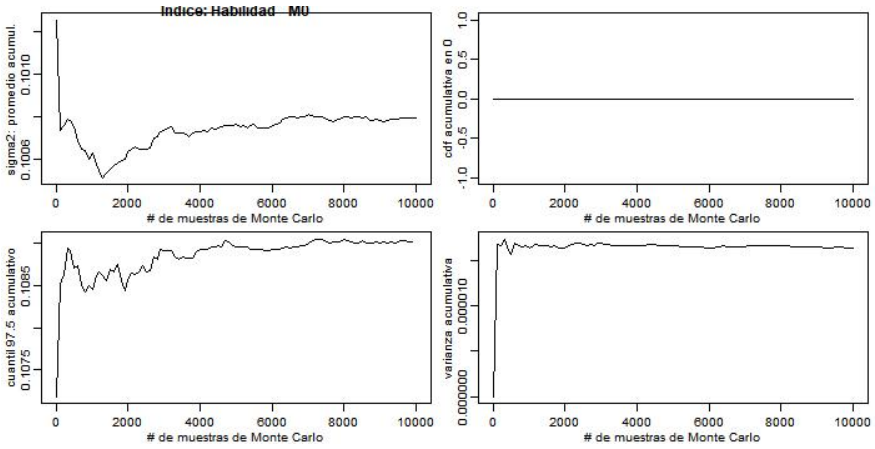
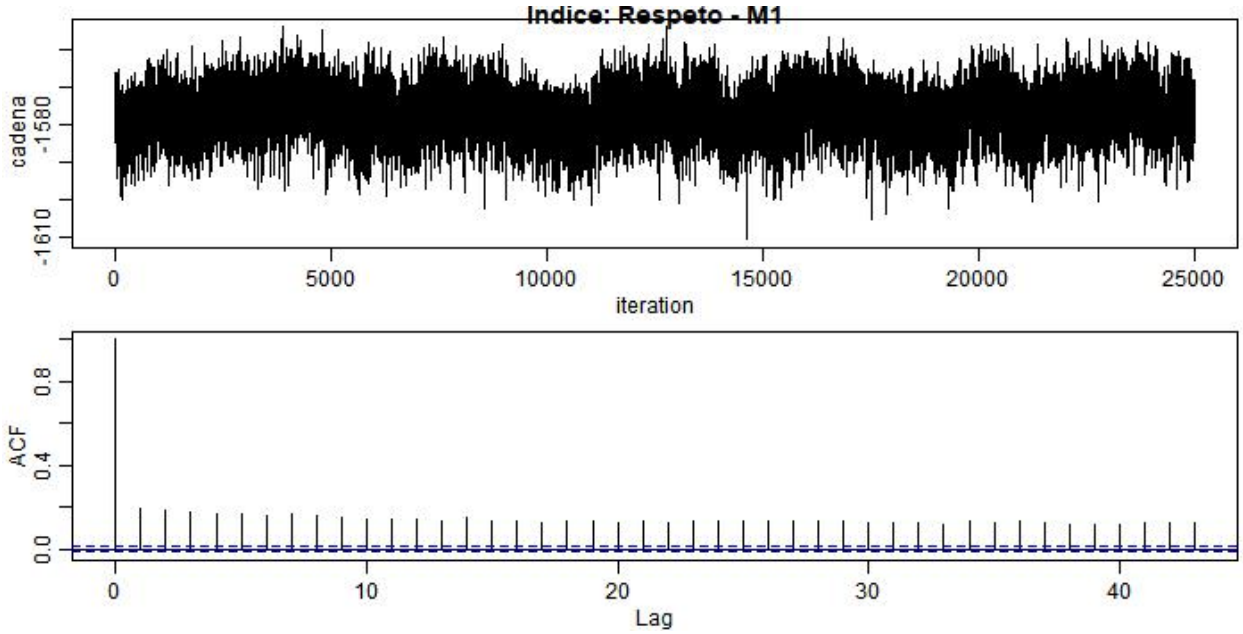



Figura 54: Convergencia del parámetro sigma2

### F.1.2 Modelo jerárquico de tres niveles




La convergencia es adecuada, pero presenta autocorrelación. Bajo pruebas formales, realizadas con el paquete *coda* de R, la cadena converge.

Obsérvese que el número de iteraciones fue de 25,000 para el modelo, como ya se explicó.

El comportamiento de las variables de interés: $\theta_{jk}$ y $\sigma_k^2$ es bueno, con muestras efectivas de 11,504 y 18,448. También los tamaños de muestra efectiva de $\mu_k$, $\gamma$ y $\kappa^2$ fueron buenos: 22,757, 25,460 y 24,999 respectivamente. Corresponde al modelamiento de la media de $\theta$. Pero los tamaños de muestra efectiva de $\tau^2$, $\lambda$ y $\xi$ fueron pésimos: 61, 82 y 41 respectivamente. Corresponden al modelamiento de la varianza de $\theta$.

En la Figura 55 se observa cómo no se logra estabilización (convergencia) del parámetro $\tau^2$, del nivel 2. Compárese con el comportamiento del parámetro $\kappa^2$ (Figura 56) del nivel 3, el cual sí presenta estabilización.

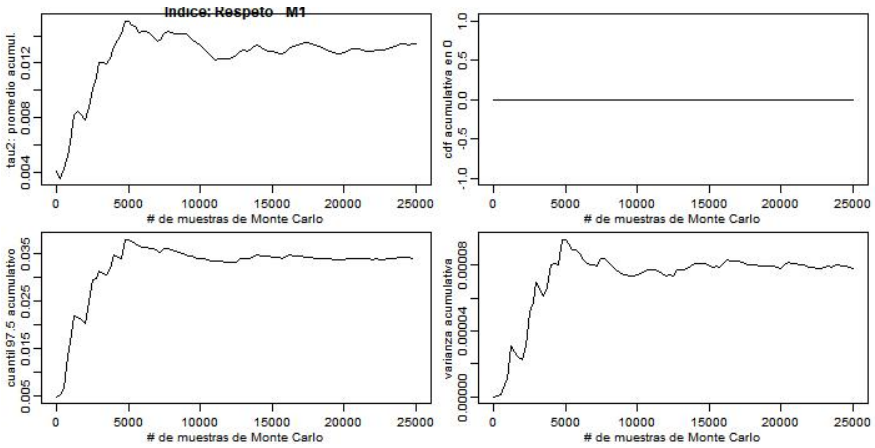



Figura 55: Convergencia del parámetro tau2

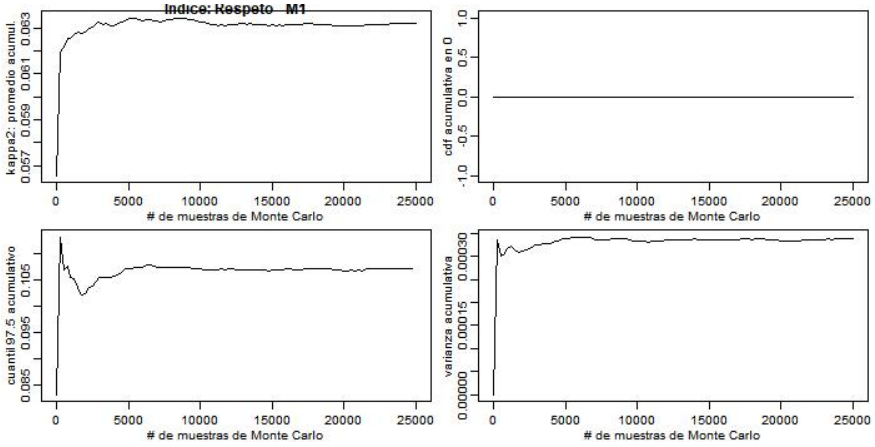



Figura 56: Convergencia del parámetro kappa2

Cuadro 25: Errores estándar del modelo jerárquico de tres niveles

| Distribución | Error estándar |
|---|---|
| Theta | 0.0011764 |
| Sigma2 | 0.0009079 |
| Mu | 0.0008178 |
| Tau2 | 0.0011330 |
| Alpha | 0.2519907 |
| Eta | 0.1713855 |
| Gamma | 0.0003438 |
| Kappa2 | 0.0001162 |
| Lambda | 26.5223759 |
| Xi | 0.8998092 |

Se observa un error estándar aparentemente alto para $\alpha$, $\eta$, $\lambda$ y $\xi$. No obstante, en relación a las medias no son valores altos (la media entre el EE es, respectivamente, 81.5, 77.7, 20.3, 7.8).

El modelo jerárquico de tres niveles obtuvo un AMSE de 0.831. Una desmejora del 1 % respecto al modelo base.

El DIC del modelo es -39,366.69.

Cuadro 26: Porcentaje de establecimientos educativos, en el modelo jerárquico de tres niveles, con un predictive posterior p-value dentro de los rangos esperados

| | Media | Mediana | Desviación estándar |
|---|---|---|---|
| Aula | 100 | 100 | 94.2 |
| Establecimiento educativo | 100 | 100 | 100.0 |

El modelo jerárquico de tres niveles obtiene un desempeño inferior, medido mediante el DIC, al modelo base. No obstante, es consistente para las tres estadísticas tanto a nivel de establecimiento educativo, como de aula.

### F.2.1 Modelo jerárquico de dos niveles, con covariables

El cálculo de la Log-verosimilitud del MCMC para los modelos jerárquicos de dos niveles es:

**A nivel de aula**

$$logP\left(\tilde{y}_{ijk} \mid \beta_{jk}^{(s)}, (\sigma_k^2)^{(s)}\right) = \Sigma_{k=1}^{M}\Sigma_{j=1}^{n_k}\Sigma_{i=1}^{n_{jk}} N(\beta_{jk}^{(s)} X_{ijk}^{T}, (\sigma_k^2)^{(s)})$$

**A nivel de establecimiento educativo**

$$logP\left(\tilde{y}_{ijk} \mid \beta_{k}^{(s)}, (\sigma_k^2)^{(s)}\right) = \Sigma_{k=1}^{M}\Sigma_{j=1}^{n_k}\Sigma_{i=1}^{n_{jk}} N(\beta_{k}^{(s)} X_{ijk}^{T}, (\sigma_k^2)^{(s)})$$

Dónde el superíndice (s) representa la iteración que corresponde.

$$(s) \in \{1, 2, 3, \ldots, S\}$$

S es el número total de iteraciones.

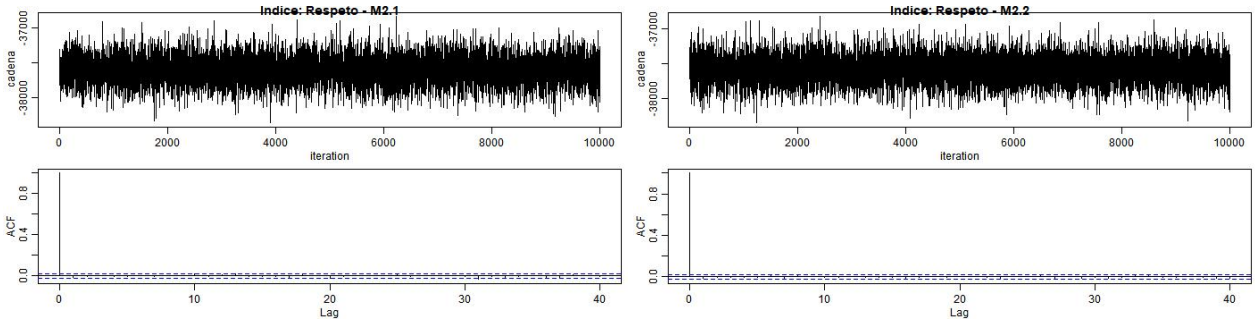



Figura 57: Convergencia de los modelos para el indice de Respeto

Las gráficas de la Figura 57 presentan la cadena de la logverosimilitud para cada uno de los modelos. Las dos opciones del modelo convergieron de manera adecuada. Para ambos se utilizó una configuración de iteraciones de MCMC: 102,500 iteraciones, con un calentamiento de la serie de 2,500 iteraciones, guardada de los resultados cada 10 iteraciones con el objeto de obtener 10,000 muestras.

Cuadro 27: Errores estándar del modelo M2.1 para el quinto establecimiento educativo

| Distribución | Error estándar |
|---|---|
| Beta_0 | 0.0004193 |
| Beta_1 | 0.0005299 |
| Sigma2 | 0.0000039 |

Los tamaños efectivos para, por ejemplo, $\beta_0$, $\beta_1$ y $\sigma^2$ del primer modelo jerárquico de dos niveles fueron: 9,953, 10,025 y 2,239,557 respectivamente.

Cuadro 28: Errores estándar del modelo M2.2 para el quinto establecimiento educativo

| Distribución | Error estándar |
|---|---|
| Beta_0 | 0.0214981 |
| Beta_1 | 0.0214863 |
| Beta_2 | 0.0376457 |
| Beta_3 | 0.0215888 |
| Sigma2 | 0.0000048 |

Los tamaños efectivos para el segundo modelo jerárquico de dos niveles fueron: 9,806, 9,817, 9,680, 9,722 y 10,039 respectivamente.

En comparación con el modelo por sexo, el modelo por grado obtiene un mayor error estándar en sus coeficientes $\beta$ (la media de los betas entre sus respectivos EE es 8.7, 23.9, 0.4, 15.6).

El primer modelo jerárquico de dos niveles obtuvo un AMSE de 0.889 en la validación cruzada. Una mejora del 5.9 % respecto al Modelo base y del 6.9 % respecto al modelo jerárquico de tres niveles.

Se presenta a continuación el ajuste asociado al Deviance Information Criterion de los modelos jerárquicos de dos niveles:

**A nivel de aula**

$$DIC = 2[-\ln \mathrm{P}(y \mid \hat{\beta}_{Bayes}) + \mathrm{P}_{\mathrm{DIC}}]$$

donde

$$\ln \mathrm{P}(y \mid \hat{\beta}_{Bayes}) = \sum_{k=1}^{M} \sum_{j=1}^{n_j} \sum_{i=1}^{n_{jk}} \ln\Big[\mathrm{N}(y_{ijk} \mid X_{ijk}^T \bar{\beta}_{jk}, \bar{\sigma}_k^2)\Big]$$

y

$$\mathrm{P}_{\mathrm{DIC}} = 2\Big(\ln \mathrm{P}(y \mid \hat{\beta}_{Bayes}) - \frac{1}{S} \sum_{s=1}^{S} \log \mathrm{P}(y \mid X_{ijk}^T \beta_{jk}^{(s)})\Big)$$

**A nivel de establecimiento educativo**

$$DIC = 2[-\ln \mathrm{P}(y \mid \hat{\beta}_{Bayes}) + \mathrm{P}_{\mathrm{DIC}}]$$

donde

$$\ln \mathrm{P}(y \mid \hat{\beta}_{Bayes}) = \sum_{k=1}^{M} \sum_{j=1}^{n_j} \sum_{i=1}^{n_{jk}} \ln\Big[\mathrm{N}(y_{ijk} \mid X_{ijk}^T \bar{\beta}_{k}, \bar{\sigma}_k^2)\Big]$$

y

$$\mathrm{P}_{\mathrm{DIC}} = 2\Big(\ln \mathrm{P}(y \mid \hat{\beta}_{Bayes}) - \frac{1}{S} \sum_{s=1}^{S} \log \mathrm{P}(y \mid X_{ijk}^T \beta_{k}^{(s)})\Big)$$

El DIC de los modelos es:

Cuadro 29: Deviance Information Criterion por modelo

| Modelo | DIC |
|---|---|
| Sexo | 4,841 |
| Grado | 6,168 |

Los modelos jerárquicos de dos niveles obtienen un DIC muy superior en valor, muy inferior en desempeño al modelo base.

A pesar del DIC mayor, los modelos jerárquicos de dos niveles permiten realizar inferencias sobre sexo y grado respectivamente, lo cual es una ganancia frente al aparente mejor desempeño de los otros dos modelos.

Cuadro 30: Porcentaje de aulas, en los modelos jerárquicos de dos niveles, con un predictive posterior p-value dentro de los rangos esperados

| | Media | Mediana | Desviación estándar |
|---|---|---|---|
| Sexo a nivel de Aula | 100 | 34.8 | 0 |
| Grado a nivel de establecimiento educativo | 100 | 46.4 | 0 |

Los modelos jerárquicos de dos niveles tienen un desempeño regular en la representación de lo que ocurre en las aulas o establecimientos educativos para las estadísticas de mediana y desviación estándar, sobre todo a nivel de desviación estándar. Pero para la media la representación es adecuada.

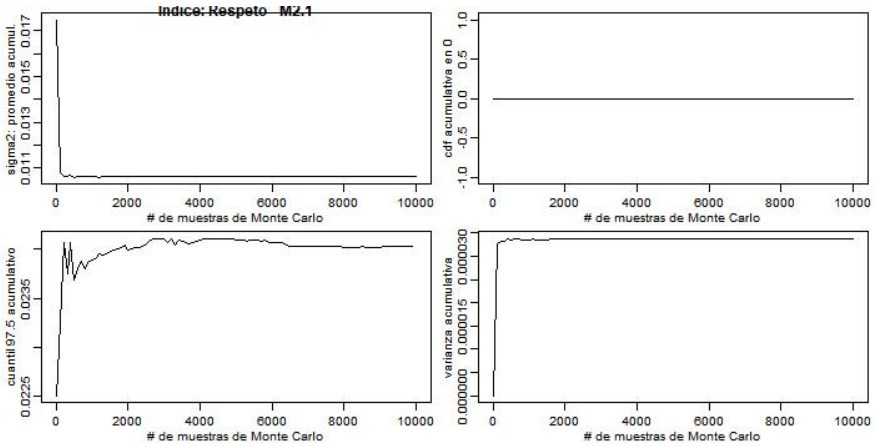



Figura 58: Convergencia del parámetro sigma2 para el primer modelo jerárquico de dos niveles

Las figuras 58 y 59 presentan que también los modelos jerárquicos de dos niveles logran convergencia del parámetro $\sigma^2$.

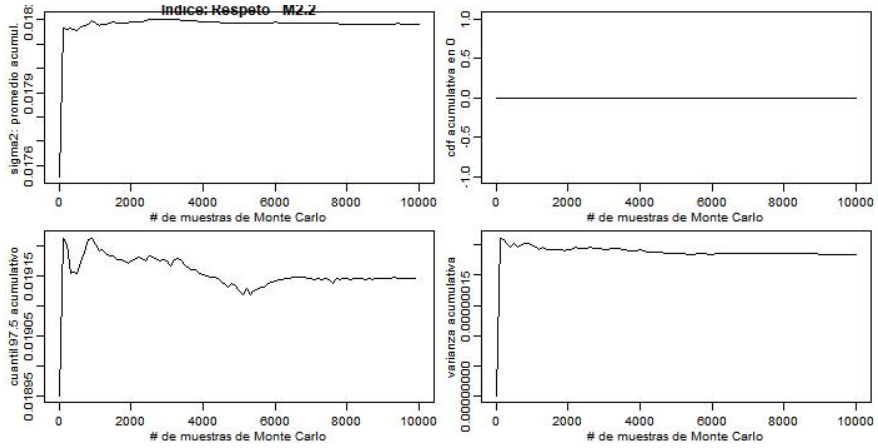



Figura 59: Convergencia del parámetro sigma2 para el segundo modelo jerárquico de dos niveles