Universidad Nacional de Educación a Distancia
Escuela Técnica Superior de Informática
Máster en Ingeniería y Ciencia de Datos

Trabajo Fin de Máster

Utilización de técnicas multivariantes para el estudio del aprendizaje de la mejora de la accesibilidad en el subtitulado de vídeos

Autor: Javier Pérez Arteaga
Directores: Emilio Letón Molina
Jorge Pérez Martín
Fecha de realización: junio 2023

Resumen

En este trabajo se ha analizado si los estudiantes de la Sexta Edición (2022) del MOOC Materiales Digitales Accesibles perteneciente al Canal Fundación ONCE en UNED son capaces de evaluar las diferencias en la calidad de subtitulado de dos vídeos. Los estudiantes que voluntariamente decidieron participar en la actividad vieron dos vídeos idénticos, uno correctamente subtitulado y otro en el que se habían introducido errores en el subtitulado. El orden en el que se presentaron los vídeos a cada estudiante fue asignado aleatoriamente. Después de ver cada vídeo, los estudiantes respondieron a una escala de Likert de dieciocho ítems con cinco niveles. El estudio fue triple ciego ya que ni los estudiantes conocían la calidad del subtitulado del vídeo que estaban viendo, ni esa información era conocida por los instructores del curso, ni lo fue a la hora de realizar el análisis estadístico. Esta información fue revelada en la fase de discusión de este trabajo. El modelado estadístico realizado tuvo en cuenta la naturaleza ordinal y longitudinal de los datos. Como variable dependiente se utilizó la respuesta a los test y como variables explicativas el nivel de subtitulado, el periodo y la secuencia de visualización, el estudiante y el ítem de Likert. Se propusieron dos tipos de modelos lineales generalizados mixtos: El primero fue una Regresión Logística en el que la variable respuesta se dicotomizó y el segundo una Regresión Ordinal Acumulativa. Con este último modelo se realizó tanto un análisis frecuentista como bayesiano. Las conclusiones de la exploración de los datos y del modelado estadístico coinciden en que estudiantes nóveles en accesibilidad fueron capaces de evaluar las diferencias de calidad en el subtitulado. Particularmente, percibieron diferencias en la corrección ortográfica y gramatical, la literalidad, la identificación de los personajes, etc. Sin embargo, tuvieron dificultades en la percepción de la calidad cuando se trata de aspectos espaciales (número de líneas y de caracteres por línea) y temporales (sincronización y velocidad) del subtitulado.

Agradecimientos

Agradezco sinceramente el interés demostrado y la dedicación de mis directores de TFM, Emilio Letón y Jorge Pérez. Valoro especialmente las innumerables horas que hemos pasado reunidos, el tiempo que habéis dedicado a revisar mi trabajo y la confianza depositada en mí. Vuestro conocimiento experto en temas de accesibilidad y en modelado estadístico y los valiosos consejos que he recibido para organizar y estructurar este documento han contribuido sin duda a hacer mejor este trabajo. Confío en que mantengamos la colaboración en el futuro y pueda seguir aprendiendo de vosotros. Muchas gracias, Jorge. Muchas gracias, Emilio.

Javier Pérez Arteaga
Madrid, junio de 2023

Índice

Índice de tablas

Índice de figuras

Glosario

accesibilidad Condición que deben cumplir los entornos, productos y servicios para que sean comprensibles, utilizables y practicables por todos los ciudadanos, incluidas las personas con discapacidad. 1, 46, 73

ANOVA El Análisis de la Varianza es un método estadístico para comparar las medias de dos o más grupos comparando la varianza de los datos entre grupos e intra grupos. xv, 9

CRC Comprobación de Redundancia Cíclica. Es un algoritmo de detección de errores utilizado para garantizar la integridad de la información. 22, 23

diseño completamente aleatorizado Diseño experimental en el cual los sujetos de estudio o las unidades experimentales se asignan de manera aleatoria a diferentes tratamientos o grupos de tratamiento. En este tipo de diseño, cada unidad experimental tiene la misma probabilidad de ser asignada a cualquier tratamiento en particular. Se utiliza cuando no hay restricciones o consideraciones específicas sobre como se deben asignar las unidades experimentales a los tratamientos. Se asume que las unidades experimentales son homogéneas y que no hay factores de confusión o variables ocultas que puedan influir en los resultados del experimento. Este diseño es útil cuando el objetivo principal del estudio es comparar los efectos de diferentes tratamientos. Al asignar las unidades de forma aleatoria, se busca reducir el sesgo y controlar los factores de confusión desconocidos, ya que se espera que las diferencias observadas entre los grupos sean principalmente el resultado de los tratamientos aplicados. El diseño completamente aleatorizado es considerado uno de los diseños experimentales más simples y robustos. Sin embargo, puede requerir un tamaño de muestra más grande para detectar diferencias significativas entre los grupos de tratamiento debido a la aleatorización pura y la posible variabilidad inherente en los datos. 7

diseño cruzado Diseño experimental en el cual los sujetos reciben diferentes tratamientos en varios momentos o periodos. Se realiza una asignación aleatoria para determinar el orden en el cual los tratamientos son recibidos. El diseño más simple de este tipo involucra dos grupos de sujetos, uno de los cuales recibe cada uno de dos tratamientos, A y B, en el orden AB, mientras que el otro los recibe en orden inverso, BA. Dado que la comparación

de tratamientos se realiza "dentro del sujeto" en lugar de "entre sujetos", es probable que se necesiten menos sujetos para lograr un poder estadístico determinado. El análisis de estos diseños no es necesariamente sencillo debido a la posibilidad de efectos de secuencia y periodo. Para minimizar este problema, a menudo se deja un tiempo de lavado entre tratamientos. xiv, 7

efecto periodo En un diseño experimental es el efecto que se produce cuando la aplicación de un tratamiento en un periodo es afectado por la aplicación de otro tratamiento en un periodo anterior. 9, 37, 57

efecto secuencia En un diseño cruzado es el efecto que se produce cuando el orden de las intervenciones afecta el resultado final. 9, 42, 46, 59, 81

error de tipo I El error de tipo I o falso positivo es el error que se comete cuando se rechaza la hipótesis nula siendo esta verdadera en la población. 9

error de tipo II El error de tipo II o falso negativo es el error que se comete cuando no se rechaza la hipótesis nula siendo esta falsa en la población. 9

escala de Likert Es una escala de evaluación ordinal utilizada en cuestionarios. Suele constar de 5 a 7 niveles desde "Totalmente en desacuerdo" hasta "Totalmente de acuerdo." 9, 22, 23, 46, 65

función logit Función matemática utilizada en el contexto de la Regresión Logística para transformar la probabilidad de un evento en un valor continuo que abarca todo el intervalo real. La función logit se define como el logaritmo natural del *odds* de un evento:

$$logit(evento) = log(\frac{P(evento)}{1 - P(evento)})$$

xiv, 11

función logística También conocida como función sigmoidea, es la función inversa de la función logit. Es una función matemática que transforma una variable continua en un rango de valores entre 0 y 1. Se utiliza comúnmente en la Regresión Logística para convertir los resultados de una combinación lineal de variables predictoras en probabilidades:

$$P(x) = \frac{1}{1 + e^{-x}}$$

11

GLM Modelo Lineal Generalizado. Es una extensión del Modelo Lineal General que permite variables de respuesta que siguen distribuciones de probabilidad no normales. xv, xvii, 10

GLMM Modelo Lineal Generalizado Mixto. Modelo estadístico también conocido como Modelo Multinivel o Modelo Jerárquico. Es una extensión de los modelos GLM utilizado para analizar datos estructurados en diferentes niveles o niveles jerárquicos. Este tipo de modelo es útil cuando los datos tienen una estructura anidada, como estudiantes dentro de escuelas, pacientes dentro de hospitales o empleados dentro de empresas. En un modelo multinivel, se reconoce que los datos están agrupados en diferentes niveles y que las observaciones dentro de cada nivel son dependientes debido a la influencia del nivel al que pertenecen. Por ejemplo, las puntuaciones de los estudiantes dentro de una escuela pueden no ser independientes debido a la influencia del entorno escolar. El modelo multinivel permite modelar tanto las variaciones entre los diferentes niveles (variaciones entre escuelas, hospitales, empresas, etc.) como las variaciones dentro de cada nivel (variaciones individuales dentro de cada escuela, hospital, empresa, etc.). Esto se logra mediante la inclusión de términos aleatorios en el modelo que capturan las variaciones entre los niveles. La ventaja de utilizar un modelo multinivel es que tiene en cuenta la estructura jerárquica de los datos y permite estimar los efectos tanto a nivel individual como a nivel de grupo. La estimación de un modelo multinivel generalmente se realiza mediante métodos de Máxima Verosimilitud Restringida (REML) o mediante el enfoque de Estimación de Máxima Verosimilitud (MLE). xv, xvii, 15

ICC La Correlación Intraclase es una medida estadística utilizada para cuantificar la proporción de varianza total de una variable que se debe a la variación entre grupos. Se utiliza en el contexto de modelos GLMM cuando las observaciones están agrupadas en diferentes unidades, y mide la parte de la varianza explicada por los efectos aleatorios. 16, 59

LRT El Test de Razón de Verosimilitudes en español, es una prueba estadística utilizada para comparar dos modelos estadísticos y determinar cuál de ellos proporciona un mejor ajuste a los datos observados. La idea detrás del LRT es comparar la verosimilitud del modelo completo con la verosimilitud del modelo reducido, que es un modelo en el que algunos de los parámetros se han igualado a cero. Si la diferencia entre estas verosimilitudes es estadísticamente significativa, se concluye que el modelo completo proporciona un mejor ajuste a los datos que el modelo reducido. 11, 60

MANOVA El Análisis de Varianza Multivariado una extensión de ANOVA que se utiliza cuando se tienen múltiples variables respuesta. 9

MCMC Los Métodos de Montecarlo basados en cadenas de Markov son métodos estadísticos utilizados para simular muestras de una distribución de probabilidad. Se basan en la construcción de una cadena de Markov, que es una secuencia de valores generados a partir de una distribución de probabilidad condicional dada la observación anterior en la cadena.

Estos valores se generan iterativamente, de manera que la distribución de probabilidad de los valores de la cadena converge a la distribución de interés. La principal ventaja de MCMC es que permite aproximar y obtener muestras representativas de una distribución de probabilidad incluso cuando dicha distribución no se puede obtener de manera analítica o es computacionalmente costosa de calcular. 18, 51

MLE La Estimación de Máxima Verosimilitud es un método estadístico de estimación paramétrica basado en la maximización de la función de verosimilitud de los datos observados. La verosimilitud es la probabilidad de observar los datos en función de los parámetros del modelo. El objetivo del MLE es encontrar los valores de los parámetros que maximizan la probabilidad de obtener los datos observados. En otras palabras, se busca encontrar los valores de los parámetros que hacen que los datos observados sean más probables bajo el modelo propuesto. En muchos casos, es más conveniente maximizar el logaritmo de la función de verosimilitud, ya que simplifica los cálculos y no afecta la ubicación de los máximos por ser la función logarítmica monótona creciente. 10

MOOC Curso en Línea Masivo y Abierto 2

odds Es la proporción entre dos probabilidades complementarias. Matemáticamente:

$$odds = \frac{P(evento)}{1 - P(evento)}$$

El *odds* puede tomar valores en el rango cero a infinito. Si el valor del *odds* es 1, significa que la probabilidad del evento y su complementario son iguales. Si el *odds* es mayor que 1, indica que la probabilidad del evento es mayor que la probabilidad de su complementario. Por el contrario, si el *odds* es menor que 1, significa que la probabilidad del evento es menor que la probabilidad de su complementario. xvi, 11, 36

odds ratio Es una medida de asociación entre dos variables dicotómicas. Matemáticamente se define como la razón de dos odds:

$$odds_{Y=1} = \frac{P(X = 1|Y = 1)}{1 - P(X = 1|Y = 1)}$$

$$odds_{Y=0} = \frac{P(X = 1|Y = 0)}{1 - P(X = 1|Y = 0)}$$

$$OR = OR_Y = OR_X = \frac{\frac{P(X=1|Y=1)}{1-P(X=1|Y=1)}}{\frac{P(X=1|Y=0)}{1-P(X=1|Y=0)}}$$

Si el OR es igual a 1, indica que no hay asociación entre las variables. Si el OR es mayor que 1, indica una mayor probabilidad del $odds_{Y=1}$. Si el OR es mayor que 1, indica una mayor probabilidad del $odds_{Y=0}$. 11, 36

OR odds ratio 14, 57, *véase* odds ratio

R Lenguaje de programación especializado en análisis estadístico y la generación de gráficos. 5

Regresión Logística Modelo estadístico en el que la variable respuesta es binaria o dicotómica. Es un tipo particular de GLM. xiv, xvii, 10, 11, 37, 57, 65, 66

Regresión Ordinal Modelo estadístico en el que la variable respuesta es ordinal. Es una extensión de la Regresión Logística. 12, 57, 66

SHA Secure Hash Algorithm. Es un algoritmo criptográfico cuyo propósito es generar un resumen único (hash) de mensaje que garantiza la integridad de los datos. 22, 23

shrinkage En el contexto de GLMM, el término ”shrinkage” (o ”encogimiento” en español) se refiere a un proceso mediante el cual los efectos estimados a nivel individual se ajustan hacia el valor promedio del grupo al que pertenecen. 16

triple ciego Diseño de experimento en el que tanto los participantes, como los investigadores encargados de administrar el tratamiento y los evaluadores de los resultados, desconocen quién está recibiendo cada nivel de tratamiento. 22

Capítulo 1

Introducción

1.1 Motivación

Algunas personas tienen problemas de accesibilidad a contenidos multimedia. Por ejemplo, personas sordas o con discapacidad auditiva, personas que no dominen el idioma o que lo estén aprendiendo, situaciones en las que el contenido audiovisual se reproduce en entornos ruidosos o en los que el silencio es necesario. Añadir subtítulos a los vídeos facilita que se superen estas dificultades ya que permiten la percepción visual de información que originalmente era sonora. Por ello, los subtítulos constituyen uno de los componentes fundamentales de la accesibilidad audiovisual (ver Pérez Martín et al. 2021). En las *Pautas de Accesibilidad para el Contenido Web, WCAG 2.1* (ver W3C 2018) se incluyen pautas y criterios que deben seguirse en el contenido web, como la obligatoriedad de tener subtítulos en los vídeos. La norma UNE 153010 (ver AENOR 2012) sobre *Subtitulado para personas sordas y personas con discapacidad auditiva* especifica requisitos y recomendaciones sobre subtitulado para facilitar la accesibilidad a los contenidos audiovisuales.

Las plataformas de compartición de vídeos, como YouTube, permiten en la actualidad la generación de subtítulos automáticos. Además, existen comunidades que se dedican a subtitular todo tipo de material multimedia. No obstante, frecuentemente los subtítulos producidos de esta manera no tienen en cuenta los criterios de calidad y de accesibilidad del subtitulado. Un subtitulado de calidad consta de información interpretativa e incluye descripción textual de efectos sonoros junto con otros elementos como la perfecta sincronización con el hablante, el tiempo de permanencia de subtítulo en la pantalla, el número de caracteres por línea y el número de líneas, la exactitud del diálogo, la ubicación, el tamaño, el contraste del subtítulo, etc. Así, Parton (2016) realizó un estudio para determinar si los subtítulos automáticos generados por YouTube cumplen las necesidades de los estudiantes universitarios sordos. En el estudio se contabilizaron un total de 525

errores en 68 minutos de video (una tasa de 7.7 errores por minuto). Adecuar los subtítulos a las normas de accesibilidad es una tarea importante y compleja que requiere dedicación y conocimiento específico y va más allá de verificar si el vídeo tiene o no subtítulos.

Este trabajo analiza datos que proceden de una actividad voluntaria sobre evaluación del subtitulado que se propuso a los estudiantes de la Sexta Edición de 2022 del **curso MOOC Materiales Digitales Accesibles** perteneciente al Canal Fundación ONCE en UNED. La Fundación ONCE, el Real Patronato sobre Discapacidad y la UNED colaboran para producir recursos educativos abiertos, gratuitos, de calidad y accesibles para todas las personas. El Canal se creó en 2016 y hasta marzo de 2023 casi 23.000 estudiantes se han inscrito en alguno de sus nueve cursos. En el año 2022 se matricularon 3.764 alumnos y aprobaron el 19,34% de los matriculados. Los cursos se realizan en formato MOOC y en ellos se ofrece formación en conocimientos y habilidades necesarios para diseñar productos, entornos, sistemas y servicios desde una perspectiva de diseño universal:

- Reconocer las necesidades relacionadas con la accesibilidad.
- Dar respuesta a estas necesidades, cada actor en la medida de sus posibilidades.
- Integrar soluciones de accesibilidad universales y específicas utilizando la tecnología apropiada.

El curso de Materiales Digitales Accesibles está dirigido por los profesores Emilio Letón Molina y Alejandro Rodríguez Ascaso y se viene realizando desde 2017. Se han matriculado en alguna de sus siete ediciones (hasta mayo de 2023) más de 8.000 alumnos y tiene un porcentaje medio de aprobados sobre matriculados del 14,6%. En la Sexta Edición se matricularon 1.261 alumnos y aprobaron 165 (13,08%). El funcionamiento del Canal se articula en varias líneas de actuación. Las más cercanas a este trabajo son las de autoría y actualización de contenidos; diseño instruccional y producción de contenidos y actividades de aprendizaje; virtualización conforme a las WCAG y a las convenciones de open edX; difusión institucional y por canales digitales; atención docente y tutorial. Además, para monitorizar y optimizar la calidad de los materiales y del diseño, desde el Canal se realizan también labores de investigación en el campo de los materiales digitales y la accesibilidad, línea de investigación en la que se enmarca el trabajo presente.

El MOOC busca la formación de los estudiantes para generar materiales digitales accesibles y para que aprendan a evaluar la accesibilidad de los mismos. Este trabajo tiene como objetivo evaluar si los estudiantes son capaces de valorar adecuadamente la calidad del subtitulado. Molanes-López et al. (2021) realizaron una experiencia similar y llegaron a la conclusión de que evaluadores novatos pueden identificar problemas de accesibilidad en vídeos. En la misma línea de investigación, la evaluación social propuesta por Takagi et al. (2008) se podría aplicar a los contenidos de vídeo en la Web. A pesar de esto, ambos estudios

constataron que los evaluadores novatos pueden pasar por alto problemas sutiles de accesibilidad que requieren un conocimiento experto como, por ejemplo, la evaluación del contraste, ya que requiere de herramientas de comprobación adecuadas.

1.2 Objetivos

A los estudiantes del MOOC se les propuso una actividad consistente en evaluar la calidad del subtitulado de dos vídeos, uno correctamente subtitulado y otro con errores (ver Sección 3.1).

El objetivo de este trabajo es responder a la siguiente pregunta de investigación:

Pregunta de investigación

¿Son los estudiantes de un curso de creación de materiales accesibles capaces de evaluar las diferencias en la calidad del subtitulado de un vídeo?

Además, también se responderá a los siguientes objetivos específicos:

Objetivo específico

¿En qué pautas de subtitulado los estudiantes tienen mayor **facilidad** para reconocer diferencias entre un subtitulado correcto y otro incorrecto?

Objetivo específico

¿En qué pautas de subtitulado los estudiantes tienen mayor **dificultad** para reconocer diferencias entre un subtitulado correcto y otro incorrecto?

Objetivo específico

¿Son los estudiantes capaces de valorar de forma similar los aspectos del subtitulado que no cambian en los vídeos?

Objetivo específico

Efecto secuencia: ¿El orden en el que vieron los vídeos los estudiantes influye en la calidad del subtitulado percibida?

Objetivo específico

Efecto periodo: ¿La evaluación del subtitulado del segundo vídeo visto está influida por haber evaluado un vídeo previamente?

1.3 Organización del trabajo

En el capítulo **Marco teórico y estado del arte** (ver Capítulo 2) se enmarca la actividad de subtitulado en el contexto del modelado estadístico, describiendo sus principales características y proponiendo y justificando las técnicas y modelos que se van a utilizar. En este capítulo también se explica la forma en que se deben interpretar y evaluar los modelos.

El capítulo **Materiales y métodos** (ver Capítulo 3) describe la actividad de subtitulado evaluada por los alumnos, los ficheros de datos suministrados, la actividad de preprocesado realizada sobre los mismos y las variables que se utilizarán en el modelado estadístico.

El capítulo **Modelado estadístico** (ver Capítulo 4) comienza con un Análisis Exploratorio de los datos, tras el que se describe como se han aplicado las técnicas de modelado presentadas en el Marco Teórico al diseño del experimento de la actividad de subtitulado.

En el capítulo de **Resultados** (ver Capítulo 5) se presentan los resultados de los modelos seleccionados en el capítulo anterior.

En el capítulo **Discusión** (ver Capítulo 6) se utilizan los resultados del capítulo anterior para responder a la pregunta de investigación y a los objetivos específicos y se plantean las limitaciones del estudio.

Finalmente, el capítulo **Conclusión y trabajo futuro** (ver Capítulo 7) se destina a recapitular los hallazgos encontrados aventurando posibles explicaciones a los mismos y propone líneas de investigación futuras en base a los resultados obtenidos.

1.4 Convenciones usadas

En este trabajo se ha evitado en la medida de lo posible el uso de anglicismos traduciendo al español los términos ingleses cuando su uso sea habitual en la publicación científica en español. No obstante, algunos términos se han mantenido en inglés por no tener una traducción fácil o frecuente. Es el caso, por ejemplo, de *odds* y de *odds ratio*. Se ha considerado que la utilización de los términos en español, *disparidad* y *razón de disparidades* respectivamente, dificultan la comprensión y se ha preferido el inglés en estos casos y otros similares. También se han introducido palabras como *frecuentista*, *bayesiano*, *dicotomizar* o *instruccional* que, aun cuando no figuren en el Diccionario de la RAE, tienen una construcción correcta en español.

Los nombres de los modelos estadísticos se han escrito con la inicial de cada palabra en mayúscula. En los acrónimos generalmente se ha mantenido su correspondencia en inglés. Por ejemplo, Modelo Lineal Generalizado (GLM, Generalized Linear Model).

Para denominar las variables utilizadas en el modelado estadístico se ha preferido

el inglés. Por ejemplo, *Treat* para referirse a los subtitulados [1] y *Response* para las respuestas a los ítems de las escalas de Likert. La justificación de esta decisión es evitar la mezcla de idiomas en los resúmenes de los modelos o en los ejemplos de código.

El trabajo se ha elaborado siguiendo las pautas de reproducibilidad recomendadas en el desarrollo de una investigación científica. Se ha realizado con la herramienta de publicación científica Quarto que integra lenguajes como el lenguaje de programación R y de publicación como Markdown o LaTeX. Todas las figuras mostradas han sido generadas con R.

[1] En investigación médica es habitual denominar tratamiento al medicamento en estudio. Por analogía, en este trabajo se llamará tratamiento al subtitulado.

Capítulo 2

Marco teórico y estado del arte

2.1 Características del diseño del experimento

En este trabajo se estudiarán las diferencias existentes entre dos niveles de subtitulado (uno correcto y otro con errores) a través de las respuestas de los estudiantes a una escala de Likert de 18 ítems que fue respondida tras visualizar cada vídeo (ver Sección 3.1). Para ello se propondrán modelos estadísticos adecuados al diseño del experimento.

El diseño del experimento de la actividad de subtitulado fue completamente aleatorizado y **cruzado *AB/BA***. En estos diseños se desea conocer el efecto de un factor con dos niveles sobre una variable respuesta. Para ello, se asigna aleatoriamente a los participantes a dos grupos y se mide la variable respuesta en dos periodos en cada grupo. En el primer periodo, a uno de los grupos se le asigna un nivel de factor y al otro grupo el otro nivel de factor. En el segundo periodo se intercambian los niveles de factor asignados a cada grupo (ver Figura 2.1). Este diseño se diferencia del diseño paralelo AA/BB en el que el nivel de factor asignado a cada grupo se mantiene entre periodos (ver Figura 2.2). El diseño cruzado y paralelo se pueden combinar si los participantes se clasifican en cuatro grupos ($AA/AB/BA/BB$). Por último, se pueden hacer diseños con más periodos o/y con más factores o niveles de factor. Los diseños cruzados son habituales en estudios clínicos en investigación médica (ver Lim e In 2021) y farmacológica para la evaluación de medicamentos genéricos.

Un **diseño completamente aleatorizado** (Lawson 2015, pp. 18) «garantiza la validez del experimento contra sesgos causados por otras variables ocultas. Cuando las unidades experimentales se asignan aleatoriamente a los niveles de factor de tratamiento, se puede realizar una prueba exacta de la hipótesis de que el efecto

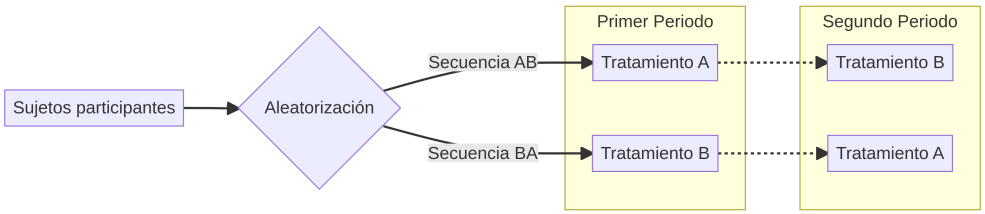

Figura 2.1: Diagrama de diseño cruzado AB/BA

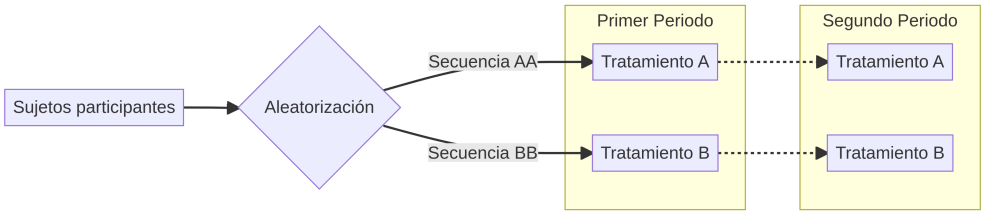

Figura 2.2: Diagrama de diseño paralelo AA/BB

del tratamiento es cero utilizando una prueba de aleatorización».

Siguiendo a Senn (2022, pp. 5-9), para que el ensayo sea de tipo cruzado no sería suficiente intercambiar las secuencias sino que debe ser objeto del ensayo el estudio de las diferencias entre los tratamientos individuales que componen las secuencias. En la misma línea, Lui (2016, pp. 1-2) afirma que «el objetivo principal de un diseño cruzado es estudiar las diferencias entre tratamientos individuales (en lugar de las diferencias entre secuencias de tratamiento). Debido a que cada paciente sirve como su propio control, el diseño cruzado es una alternativa útil al diseño de grupos paralelos para aumentar la potencia».

Los principales problemas de un diseño cruzado son el abandono (`drop-out`), de alguno de los participantes y la interacción entre el tratamiento y el periodo o efecto secuencia (`carry-over` o contaminación). Además, el análisis estadístico es más complicado, particularmente cuando la respuesta es ordinal y hay más de dos tratamientos. Aplicado al experimento del subtitulado, se producirá **efecto secuencia** si las respuestas a los cuestionarios fueran diferentes cuando los vídeos se ven en un orden que cuando se ven en el otro. Además hay que tener en consideración de la existencia del **efecto periodo**, que se producirá si las respuestas del segundo periodo están influidas por haber realizado la primera actividad de subtitulado. Como ejemplo de estos efectos se puede ver Senn (2022, pp. 35-53): Se analiza un experimento que consistió en medir el flujo espiratorio máximo (PEF, por sus siglas en inglés) en 13 niños con edades entre los 7 y los 14 años con asma a los que se les administró salbutamol (un conocido broncodilatador) y formoterol (un broncodilatador de reciente aparición en el

momento en que se realizó el estudio). Se hicieron dos grupos a los que se les administró ambos tratamientos en orden inverso dejando un periodo de lavado (`washout period`) entre aplicaciones. El efecto periodo resulta de medir si el PEF medio de los dos grupos es diferente entre periodos, y el efecto secuencia consiste en comprobar si hay diferencias significativas entre aplicar primero el tratamiento con salbutamol y luego con formoterol o hacerlo al revés.

Otra cuestión de relevancia es que las respuestas a los ítems de una escala de Likert son de **tipo ordinal**. Los test estadísticos *ANOVA* o *MANOVA* presuponen que la variable de respuesta es cuantitativa y con distribución normal. Tratar las respuestas a una escala de Likert como si fueran cuantitativas no es correcto por las siguientes razones:

- Los niveles de respuesta no son necesariamente equidistantes: la distancia entre cada par de opciones de respuesta correlativos puede no ser la misma para todos los pares. Por ejemplo, la diferencia entre «Muy en desacuerdo» y «En desacuerdo» y la diferencia entre «De acuerdo» y «Muy de acuerdo» es de un nivel, pero psicológicamente puede ser percibida de forma diferente por cada sujeto.
- La distribución de las respuestas ordinales puede ser no normal. En particular esto sucederá si hay muchas respuestas en los extremos del cuestionario.
- Las varianzas de las variables no observadas que subyacen a las variables ordinales observadas pueden diferir entre grupos, tratamientos, periodos, etc.

En Liddell y Kruschke (2018) se han analizado los problemas potenciales de tratar datos ordinales como si fueran cuantitativos constatando que se pueden presentar las siguientes situaciones:

- Se pueden encontrar diferencias significativas entre grupos cuando no las hay: error de tipo I.
- Se pueden obviar diferencias cuando en realidad sí existen: error de tipo II.
- Incluso se pueden invertir los efectos de un tratamiento.
- También puede malinterpretarse la interacción entre factores.

Otra cuestión que hay que tener en cuenta es que, al tratarse de un diseño cruzado, es de **medidas repetidas** ya que cada sujeto realiza dos veces el test, uno con cada vídeo y que, por lo tanto, las respuestas a cada test de un mismo sujeto no son independientes. Además, tampoco se pueden considerar independientes los ítems que componen el test ya que los ítems pretenden medir la misma variable latente: la calidad del subtitulado.

En este trabajo se analiza si el nivel de subtitulado (correcto o defectuoso) influye en el nivel de respuesta a los ítems de la escala de Likert, que es la variable

dependiente. Se evalúa también la existencia de efectos secuencia y periodo y la influencia que tienen sobre el nivel de respuesta el estudiante y el propio ítem.

2.2 Modelos Lineales Generalizados [1]

El Modelo Lineal Generalizado (*Generalized Linear Model*, *GLM*) es un modelo en el que la variable respuesta no sigue una distribución Normal. Para especificar un GLM son necesarios tres componentes (ver Agresti 2018, pp. 66-67):

- Un componente aleatorio que será una distribución de probabilidad de la familia exponencial. Se asume que la variable respuesta Y se distribuye según este componente aleatorio.
- Un componente lineal de predictores:

$$\tau + \beta_1 x_1 + ... + \beta_p x_p$$

- Una función de enlace g que relaciona $\mu = E(Y)$ con los predictores, de tal forma que:

$$g(\mu) = \tau + \beta_1 x_1 + ... + \beta_p x_p$$

La estimación de coeficientes en *GLM* se realiza maximizando la función de verosimilitud (*Maximum Likelihood Estimation*, *MLE*). Es decir, que los coeficientes del modelo son aquellos que maximizan la probabilidad de los datos.

Regresión Logística

La Regresión Logística es un caso particular de *GLM* en el que la variable respuesta es dicotómica. Aunque la Regresión Logística no es aplicable directamente a las respuestas de una escala de Likert por ser éstas ordinales, se introduce aquí por dos motivos:

- En la sección de modelado (ver Sección 4.2), se propondrán dos transformaciones de la variable respuesta para convertirla en dicotómica.
- Además, en este capítulo se introducirá la Regresión Ordinal (ver Sección 2.2). Este modelo se puede considerar una extensión de la Regresión Logística y permitirá tratar la variable respuesta como ordinal. De ahí el interés en presentar previamente la Regresión Logística.

[1] No se deben confundir los Modelos Lineales Generales con los Modelos Lineales Generalizados. En los primeros, también llamados Modelos de Regresión Multivariante, se presupone que las variables respuesta tienen una relación lineal con los predictores y sus valores se distribuyen normalmente. Los segundos son una generalización de los primeros y permiten que la variable respuesta admita otras distribuciones además de la normal.

Como se ha dicho, la Regresión Logística (ver Agresti 2018, pp. 68-69) es un caso particular de *GLM* donde la variable respuesta, Y, es dicotómica o Bernoulli. Es decir, que Y toma valores 0 ó 1. En una función de Bernoulli de parámetro π ($E[Y] = P(Y = 1) = \pi$), es necesaria una función que mapee los valores que puede tomar el componente lineal de rango $(-\infty, +\infty)$ a los valores que puede tomar π en el rango $(0, 1)$. Una función que permite hacer esto es la *función logit*:

$$logit(Y = 1) = log\left[\frac{P(Y = 1)}{1 - P(Y = 1)}\right] = \tau + \beta_1 x_1 + ... + \beta_p x_p \tag{2.1}$$

La inversa de la función *logit* es la *función logística* y permite realizar el mapeo inverso para obtener la probabilidad:

$$P(Y = 1) = \frac{1}{1 + exp^{-\tau - \beta_1 x_1 ... - \beta_p x_p}}$$

La interpretación de los coeficientes es la siguiente (ver Friendly et al. 2015, p. 260):

- τ es el logaritmo del *odds* de Y cuando $x_j = 0, \forall j \in 1...p$.
- β_j es el logaritmo del *odds ratio* asociado a una unidad de incremento de x_j.

El contraste de hipótesis para los coeficientes β:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

se puede realizar con el Test de Wald:

$$W = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)} \sim N(0, 1)$$

o con el Test de Razón de Verosimilitudes (*Likelihood Ratio Test*, *LRT*):

$$\begin{aligned} LRT = \Lambda &= -2 \log \frac{L(\widehat{reducido})}{L(\widehat{completo})} \\ &= -2 \log L(\widehat{reducido}) + 2 \log L(\widehat{completo}) \sim \chi^2_r \end{aligned}$$

donde:

- r es el número de $\beta's$ iguales a cero.
- $L(\widehat{reducido})$ es el valor que maximiza la función de verosimilitud en la que algunos (r) de los $\beta's$ han sido igualados a cero.
- $L(\widehat{completo})$ es el valor que maximiza la función de verosimilitud en el modelo que incluye todos los $\beta's$.

LTR permite comprobar la hipótesis de que uno o varios coeficientes sean cero. Para comparar modelos no anidados, se puede usar el Criterio de Información de Akaike (AIC) o el Criterio de Información Bayesiano (BIC), que se definen respectivamente:

$$\begin{aligned} AIC &: -2\log L + 2p \\ BIC &: -2\log L + p\log(n) \end{aligned} \tag{2.2}$$

donde L es el valor de maxima verosimilitud y el segundo sumando es una penalización que será mayor cuanto más complejo sea el modelo ($p = número\ de\ parámetros$, $n = observaciones$).

Regresión Ordinal

Las respuestas a los ítems de una escala de Likert son ordinales. La Regresión Ordinal es una clase de *GLM* que comparte muchas similitudes con la Regresión Logística (ver Sección 2.2) pero que tiene en consideración que los valores de la variable de respuesta están ordenados [2]. Según Bürkner y Vuorre (2019, pp. 3-11) hay tres clases de Regresión Ordinal:

- Regresión Ordinal Acumulativa.
- Regresión Ordinal Secuencial.
- Regresión Ordinal Adyacente.

Las regresiones ordinales secuencial y adyacente presuponen que para alcanzar un nivel se ha tenido que pasar previamente por los anteriores. En un ítem de Likert esto carece de sentido y, por lo tanto, se descartan estos modelos y se prefiere el *Modelo Acumulativo* (*Cumulative Model*, *CM*) que además es el más utilizado (ver Bürkner y Vuorre 2019, pp. 23-24).

CM presupone que la variable ordinal observada, Y, proviene de la categorización de una variable latente (no observada) continua, $\tilde{Y}$. Hay K umbrales τ_k que particionan $\tilde{Y}$ en $K+1$ categorías ordenadas observables (ver Figura 2.3). Si se asume que $\tilde{Y}$ tiene una cierta distribución (por ejemplo, normal) con distribución acumulada F, se calcula la probabilidad de que Y sea la categoría k de esta forma:

$$Pr(Y = k) = F(\tau_k) - F(\tau_{k-1})$$

Por ejemplo en la Figura 2.3: $Pr(Y = 2) = F(\tau_2) - F(\tau_1)$. Suponiendo que $\tilde{Y}$ tenga una relación lineal los predictores:

[2] Otras variantes de la Regresión Logística son la Regresión Categórica y la Regresión Multinomial. En estos tipos de *GLM* la variable respuesta puede adoptar varios valores pero no se asume que estén ordenados. La Regresión Categórica y la Regresión Multinomial están relacionadas en el mismo sentido en que lo están la Regresión Logística con función de enlace Bernoulli y con función de enlace Binomial. Es decir, que la Regresión Categórica se usa cuando las observaciones no están agrupadas y la Multinomial cuando sí lo están.

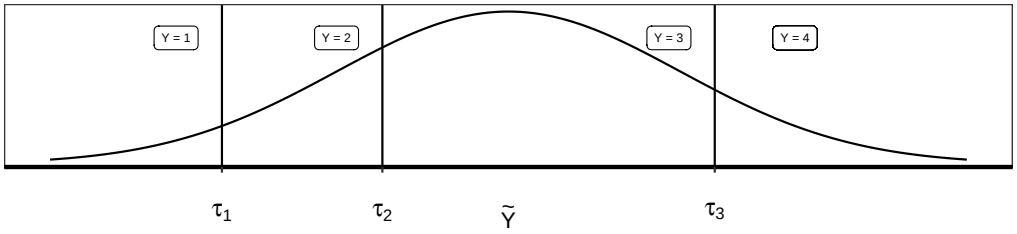

Figura 2.3: Función latente en una regresión ordinal acumulativa.

$$\tilde{Y} = \eta + \epsilon = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$$

entonces la función de probabilidad acumulada de los errores tendrá la misma forma que la de $\tilde{Y}$:

$$P(\epsilon \leq z) = F(z)$$

Se puede calcular la distribución de probabilidad acumulada de Y:

$$P(Y \leq k \mid \eta) = P(\tilde{Y} \leq \tau_k \mid \eta) = P(\eta + \epsilon \leq \tau_k) = P(\epsilon \leq \tau_k - \eta) = F(\tau_k - \eta)$$

Por lo que asumiendo la normalidad de los errores:

$$P(Y = k) = \Phi(\tau_k - \eta) - \Phi(\tau_{k-1} - \eta)$$

donde hay que estimar los umbrales τ_k y las pendientes de cada variable explicativa. La función anterior es la conocida como la función de enlace `probit`. La interpretación de los coeficientes con esta función de enlace no resulta intuitiva. Por ello en este trabajo se va a utilizar la función de enlace `logit`. Con esta función de enlace la interpretación de los coeficientes es parecida a la de los coeficientes de la regresión logística. Además, en la práctica, los coeficientes estimados suelen tener valores similares a los de la función `probit`. Para entender como se deben interpretar los coeficientes del modelo *CM* se parte del supuesto de que el *logit* de la función de probabilidad es lineal:

$$logit[P(Y \leq k)] = \tau_k - \eta = \tau_k - (\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p) \tag{2.3}$$

En ese caso, se puede demostrar fácilmente que, por ejemplo:

$$\frac{\frac{P(Y \leq k|\eta)}{P(Y > k|\eta)}}{\frac{P(Y \leq k+1|\eta)}{P(Y > k+1|\eta)}} = \exp(\tau_k - \tau_{k+1})$$

Y que [3]:

[3] En la Sección 4.2 se demuestra esta fórmula.

$$\frac{\frac{P(Y\leq k|x_j=1)}{P(Y>k|x_j=1)}}{\frac{P(Y\leq k|x_j=0)}{P(Y>k|x_j=0)}} = \exp(-\beta_j)$$

o, equivalentemente:

$$\frac{\frac{P(Y>k|x_j=x+1)}{P(Y\leq k|x_j=x+1)}}{\frac{P(Y>k|x_j=x)}{P(Y\leq k|x_j=x)}} = \exp(\beta_j)$$

Es decir, que $\exp(\beta_j)$ es el *odds ratio* (cambio relativo entre *odds*, OR) de que la variable respuesta esté por encima de una determinada categoría versus estar por debajo de ella para una unidad de incremento del predictor x_j. Un valor del coeficiente β_j positivo indica que la relación entre el predictor x_j y la función de *logit* es positiva y, por lo tanto, se incrementa la probabilidad de un mayor valor de la variable respuesta.

Presunciones del modelo

Este modelo se denomina proporcional ya que se asume que cada predictor tiene los mismos efectos sobre todos los niveles de la variable de respuesta ordinal (ver Liu 2022, chap. 5). Es decir, que los *odds* de los niveles de respuesta deben ser proporcionales para los mismos valores de las variables explicativas. Esta suposición frecuentemente no es realista y se puede relajar permitiendo estimar un coeficiente diferente para cada nivel de la variable respuesta. Sin embargo, el incremento del número de coeficientes dificulta la interpretabilidad del modelo. Harrell (2020) aboga por usar este modelo incluso aunque la suposición de proporcionalidad no se cumpla:

> «Ningún modelo se ajusta perfectamente a los datos, …, la aproximación ofrecida por el modelo *CM* sigue siendo bastante útil. Y un análisis unificado del modelo *CM* es decididamente mejor que recurrir a análisis ineficientes y arbitrarios de valores dicotomizados de Y.»

Matemáticamente la presunción de la proporcionalidad de los *odds* se demuestra a partir de la Ecuación 2.3. Si se fijan los predictores en un valor arbitrario $X = x$ y se consideran dos niveles de respuesta cualesquiera k y l, entonces:

$$\begin{aligned} logit[P(Y \leq k|X = x)] - logit[P(Y \leq l|X = x)] &= \tau_k - \tau_l \\ \frac{odds(P(Y \leq k|X = x))}{odds(P(Y \leq l|X = x))} &= \exp(\tau_k - \tau_l) \implies \\ odds(P(Y \leq k|X = x)) &\propto odds(P(Y \leq l|X = x)) \end{aligned} \tag{2.4}$$

Es decir, que la proporcionalidad de *odds* de dos niveles de respuesta es independiente de los valores concretos de los predictores, por lo que la constante de proporcionalidad debe ser similar para todos ellos.

2.3 Modelos Multinivel Generalizados

Un Modelo Multinivel Generalizado (*GLMM*, *Generalized Linear Mixed Model*), anidado, jerárquico o mixto es un modelo en el que los datos están anidados en una estructura jerárquica. Se utilizan cuando se incumple la hipótesis de independencia entre las observaciones. Por ejemplo, si se quisiera evaluar el rendimiento de varios métodos de enseñanza, se podrían seleccionar aleatoriamente varios colegios participantes y en cada uno de ellos elegir varias clases en las que se impartiría uno de los métodos de enseñanza. En este caso, los alumnos de una clase no son independientes de los alumnos de otra clase del mismo colegio y también es esperable que los alumnos de un mismo colegio sean más parecidos entre sí que los de otro colegio. Otra situación en la que se viola la condición de independencia entre observaciones es cuando se toman varias medidas del mismo sujeto. Este tipo de experimentos se llaman de medidas repetidas o longitudinales [4]. Cuando se da este supuesto, se considera que las medidas están anidadas en el sujeto (ver Liu 2022). En un modelo multinivel no es necesario que todas las variables tengan una estructura jerárquica. Se distinguen entonces dos tipos de variables: Las conocidas como de efectos fijos son aquellas que se considera que tienen el mismo efecto en toda la población y, por lo tanto, se debe estimar un único coeficiente. Las variables de efectos aleatorios tienen un coeficiente diferente para cada elemento de la población y se supone que son una muestra de una población mucho mayor, como el caso de seleccionar aleatoriamente una muestra de colegios. Normalmente el coeficiente particular de cada elemento no es de interés para el investigador y se asume que tienen una media centrada en cero. El mayor interés de los efectos aleatorios es la estimación de su matriz de varianzas-covarianzas.

La ecuación general de un modelo multinivel con dos niveles y un solo predictor con efectos aleatorios es (ver D.-G. Chen y J. Chen 2021, pp. 40):

$$\begin{aligned} Nivel\ 1: y_{ij} &= \beta_{0j} + \beta_{1j} x_{1ij} + \epsilon_{ij} \\ Nivel\ 2: \beta_{0j} &= \beta_0 + U_{0j} && (intercepto\ aleatorio) \\ \beta_{1j} &= \beta_1 + U_{1j} && (pendiente\ aleatoria) \end{aligned}$$

Los errores del modelo se distribuyen:

$$\text{Error intra grupo: } \epsilon_{ij} \sim N(0, \sigma^2)$$

$$\text{Error entre grupos: } \begin{pmatrix} U_{0j} \\ U_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \tau_0 \tau_1 \rho_{01} \\ \tau_0 \tau_1 \rho_{01} & \tau_1^2 \end{pmatrix} \right)$$

donde j son los grupos que varían $j = 1, ..., J$ (J es el número de grupos); ij es la observación i-ésima del grupo j ($i = 1, ..., n_j$, n_j es el número de observaciones

[4] Hay una diferencia conceptual entre medidas repetidas y longitudinales. Una variable se dice que es longitudinal cuando se toman varias medidas de los sujetos objeto del estudio en diferentes momentos del tiempo. Para que sea considerada de medidas repetidas, las medidas de cada sujeto se toman con distintos niveles de factor. En la práctica la distinción es poco relevante ya que ambas situaciones se parametrizan de la misma forma.

del grupo j). El modelo se compone de una parte fija $\beta_0 + \beta_1 x_{1ij}$ y una aleatoria $U_{0j} + U_{1j} x_{1ij} + \epsilon ij$. Los parámetros de este modelo son el intercepto y la pendiente de efectos fijos (β_0 y β_1), la varianza intra-grupos (σ^2), la varianza inter-grupos del intercepto aleatoria (τ_0) y de la pendiente aleatoria (τ_1), y la correlación entre intercepto y pendiente aleatorias (ρ_{01}).

En Gelman et al. (2013, p. 115) se evalúan tres posibilidades a la hora de definir un modelo:

- *Complete pooling*: Consiste en estimar un único parámetro para cada predictor. Es equivalente a un modelo con efectos fijos.
- *No pooling*: Se estiman tantos parámetros como grupos haya de forma independiente.
- *Partial pooling*: Es el modelo jerárquico. Es una mezcla de ambos, ya que, aunque se estima un parámetro para cada grupo (como en *no pooling*), esta estimación no es independiente, sino que se supone que las observaciones de un mismo grupo proceden de una misma distribución de probabilidad. Esto se traduce en que se produce una contracción (*shrinkage*) en la estimación de los parámetros hacia la media. Al influir la estimación de unas observaciones en otras, la estimación es de menor valor absoluto que la que resultaría en un modelo de *no pooling*. De esta forma se puede ver el *complete pooling* y el *no pooling* como dos casos particulares y extremos del *partial pooling*. La contracción de coeficientes en los modelos multinivel actúa como una regularización que puede evitar el sobreajuste.

Los modelos multinivel requieren supuestos adicionales en el nivel segundo y superiores que son similares a los supuestos para los modelos de efectos fijos (ver D.-G. Chen y J. Chen 2021, p. 43). Para estimar los parámetros en un modelo multinivel se suele utilizar el método de Máxima Verosimilitud Restringida ($RMLE$), que es una variante de la estimación por Máxima Verosimilitud (MLE) en la que se hacen ajustes en los grados de libertad del modelo con efectos aleatorios para corregir el sesgo que se produce al usar MLE en estos modelos.

Para evaluar si la estructura anidada es adecuada se utiliza la Correlación Intra-Clase (*ICC*, Intra-Class Correlation). ICC se puede interpretar como la proporción de la varianza explicada por la estructura de agrupamiento de la población. Se diferencia del Coeficiente de Determinación ($R2$) en que éste es la proporción de la varianza explicada por el modelo completo, mientras que ICC es la varianza explicada por los efectos aleatorios (ver Lüdecke et al. 2021). La ICC se calcula como el ratio de la varianza entre grupos y la varianza total (ver D.-G. Chen y J. Chen 2021, pp. 29-33):

$$ICC = \rho = \frac{\tau^2}{\tau^2 + \sigma^2}$$

donde τ^2 es la varianza poblacional entre grupos y σ^2 es la varianza de la población dentro del grupo. ICC oscila entre 0 (ausencia de varianza entre grupos) y 1 (no varianza intra-grupos). Cuanto más proxima a 1 sea ICC mayor evidencia

hay de la existencia de una estructura anidada. Sin embargo, no hay un consenso sobre el umbral concreto que debe tener ICC para decidir si es preferible o no la estructura anidada (ver D.-G. Chen y J. Chen 2021, p. 33). ICC se estima a partir de la varianza de los coeficientes y residual aleatorias calculadas por el modelo. En modelos *GLM* no se calcula la varianza residual y se recurre a métodos de simulación para estimar ICC.

2.4 Modelado bayesiano

El paradigma frecuentista parte de la suposición de que los datos son generados a partir de una variable aleatoria Y y para estimar los coeficientes se maximiza la función de verosimilitud $p(y|\theta)$ que depende del parámetro desconocido θ. En el análisis bayesiano se considera que θ es una variable aleatoria ya que hay incertidumbre respecto a su valor. Esto se traduce en que se debe asignar una distribución de probabilidad $p(\theta)$, conocida como distribución a priori, que expresa nuestra creencia sobre los valores que puede tomar θ. En la inferencia bayesiana se usa la distribución de probabilidad a posteriori $p(\theta|y)$ que es proporcional al producto de la función de verosimilitud y de la distribución de probabilidad a priori (ver Nicenboim Bruno 2023):

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal Likelihood}} \Rightarrow p(\theta|y) = \frac{p(y|\theta) \times p(\theta)}{p(y)} \propto p(y|\theta) \times p(\theta)$$

En la inferencia bayesiana hay dos fuentes de incertidumbre: Por un lado hay que contar con la variabilidad de Y, ya que si se toman varias muestras, los valores y_i obtenidos serán diferentes. Además, existe otra incertidumbre que proviene del desconocimiento del valor de θ. En la estimación frecuentista, debido a que se utilizan estimaciones puntuales de θ, no se tiene en cuenta esta incertidumbre. La Ecuación 2.5 se corresponde con la distribución predictiva a posteriori que tiene en consideración ambas incertidumbres:

$$\begin{aligned} p(y_{pred} \mid y) &= \int_\theta p(y_{pred}, \theta \mid y)\, d\theta = \int_\theta p(y_{pred} \mid \theta, y) p(\theta \mid y)\, d\theta \\ &= \int_\theta p(y_{pred} \mid \theta) p(\theta \mid y)\, d\theta \end{aligned} \tag{2.5}$$

donde la última igualdad resulta de la independencia condicional de y_{pred} e y dado θ ($y_{pred} \perp\!\!\!\perp y \mid \theta$).

Una crítica habitual a la inferencia bayesiana es que la elección de la distribución de probabilidad a priori es subjetiva. Aunque es cierto que hay un grado de subjetividad en esta elección, en realidad en el modelado frecuentista hay que tomar ciertas decisiones que también lo son, como por ejemplo la elección del nivel de significación o la forma que adopta la función de verosimilitud. En la práctica, si las observaciones son suficientemente informativas y la distribución a priori es poco informativa, la distribución de probabilidad a priori tendrá poca o nula influencia en la distribución a posteriori ya que estará dominada por la

función de verosimilitud y los coeficientes estimados serán muy parecidos en ambos paradigmas. Sin embargo, en lo que diferirán es en la interpretación ya que, por ejemplo, en un modelo bayesiano se pueden interpretar los intervalos de confianza como la probabilidad de que el parámetro esté dentro del intervalo. Por eso a estos intervalos se les conoce como intervalos de credibilidad. Esa interpretación en un modelo frecuentista carecería de sentido ya que los parámetros del modelo no se consideran variables aleatorias y, por lo tanto, tendrán probabilidad 1 si el verdadero valor del parámetro cae dentro del intervalo y 0 si no lo hace. Para obtener la distribución de probabilidad a posteriori normalmente se recurre a métodos del simulación *MCMC* (Métodos de Montecarlo basados en Cadenas de Markov). [5].

Para comparar modelos entre sí se pueden usar varias medidas (ver Barreda S. 2023). Por ejemplo, la conocida como *log pointwise predictive density* o densidad predictiva puntal (*lpd*) se calcula como:

$$\widehat{\text{lpd}} = \sum_{i=1}^{N} \log(p(y_i|\theta))$$

La *lpd* es la densidad conjunta de observar los datos dada la estructura del modelo y las estimaciones de los parámetros θ. Aunque las probabilidades a priori no se incluyen en su cálculo, sí influyen en la estimación de θ y, por lo tanto, tienen un efecto en los valores de *lpd*. Mayores valores de *lpd* estarían indicando un mejor modelo. El problema con esta métrica es que se utilizan los datos tanto para estimar el modelo como para seleccionar el mejor modelo. Esto va a producir un sobreajuste y tenderá a favorecer los modelos más complejos. Una métrica mejor es la *expected log pointwise predictive density* o densidad predictiva puntual esperada (*elpd*). Se define en términos de valores fuera de la muestra $\tilde{y}$ en lugar de con los valores de la muestra y:

$$\text{elpd} = \sum_{i=1}^{N} \mathbb{E}(\log(p(\tilde{y}_i|\theta)))$$

En la práctica no se puede saber el valor de *elpd* ya que no se conoce el proceso que genera verdaderos valores $\tilde{y}$. Una forma de estimar *elpd* que empíricamente se ha demostrado que funciona es penalizar *lpd* con el número de parámetros p de forma análoga a lo que se hace en AIC (ver Ecuación 2.2):

$$\widehat{\text{elpd}} = \widehat{\text{lpd}} - \text{p}$$

El problema es que en modelos multinivel conocer el número de parámetros no es sencillo ya que los parámetros asociados a efectos aleatorios no se pue-

[5]En ocasiones se puede obtener una forma análitica de la distribución a posteriori si se elige una adecuada combinación de función de verosimilitud y distribución a priori conocidas como distribuciones conjugadas. Aunque esto evita la utilización de métodos de simulación, restringe las formas posibles de las distribuciones. En la actualidad, con el aumento de la capacidad de cálculo de los ordenadores, normalmente no es necesaria la utilización de distribuciones conjugadas.

den considerar que sean completamente independientes. El número efectivo de parámetros va a depender de la importancia de la regresión hacia la media que sufra cada parámetro. Además, en lugar de usar una estimación puntual, se puede utilizar toda la distribución de valores de la simulación. La métrica *widely available information criterion* o «criterio de información ampliamente disponible» ($WAIC$) es una forma de estimar *lpd* que usa toda la distribución de probabilidad a posteriori:

$$\widehat{\text{lpd}} = \sum_{i=1}^{n} \log(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^s))$$

donde S es el tamaño de la muestra y el sumatorio interior es la media de densidad en un punto i. Para penalizar los modelos más complejos, se usa la varianza de la función de densidad logarítmica:

$$\widehat{\text{elpd}}_{WAIC} = \widehat{\text{lpd}} - \text{p}_{\text{WAIC}}$$
$$\text{p}_{\text{WAIC}} = \sum_{i=1}^{n} \text{Var}_{s=1}^{S}(\log(p(y_i|\theta^s)))$$

Una forma alternativa de evaluar un modelo es mediante validación cruzada (CV, Cross Validation). La validación cruzada más popular es *K-fold-CV*. Con esta técnica se divide el conjunto de datos en K partes y se entrena el modelo separando sucesivamente cada una de las partes que se usan para evaluar el modelo. Los valores estimados resultan de promediar los K resultados. El problema es que cuanto menor sea K más inestables son los resultados obtenidos, y cuanto mayor sea K más veces hay que reentrenar el modelo. Un caso extremo de validación cruzada, que produce gran estabilidad en las estimaciones, es dejar un dato fuera cada vez (*Leave One Out*, LOO, $K = N$). La dificultad es que requiere estimar el modelo tantas veces como datos se tengan. Para evitar esto, hay formas de aproximar $\widehat{\text{elpd}}$ basadas en LOO sin tener que reentrenar el modelo. La fórmula es la siguiente:

$$\widehat{\text{elpd}}_{LOO} \approx \sum_{i=1}^{n} \log(p(y_i|\theta_{y_{-i}}))$$

donde $\theta_{y_{-i}}$ es la estimación de θ que resulta tras eliminar la observación y_i (ver Gelman et al. 2013, pp. 175-176).

Capítulo 3

Materiales y métodos

En este capítulo se describe la actividad de subtitulado de la que proceden los datos y los ficheros de datos suministrados; se explica la tarea de preprocesado realizada sobre ellos y las variables que se han utilizado en el modelado estadístico.

3.1 Descripción de la experiencia

La actividad de subtitulado, cuyos resultados se analizan en este trabajo, ha sido cuidadosamente diseñada con un enfoque instruccional curado a través de las sucesivas ediciones del curso y así ofrecer los contenidos de forma alineada para que los estudiantes puedan realizar la evaluación del subtitulado de forma secuencial.

La actividad fue voluntaria y sin influencia en la calificación final del alumno. Se realizó en el módulo «Accesibilidad del material multimedia». En este mismo módulo, y antes de la actividad de subtitulado, los alumnos hubieron completado las secciones «Accesibilidad de la información sonora» y «Accesibilidad de la información visual». Estos módulos constan de las siguientes actividades relacionadas con la actividad de subtitulado:

- Vídeo: «Accesibilidad audiovisual: subtítulos».
- Invitación a la participación en el foro.
- Textos con referencias.
- Test de nivel 1 y nivel 2: «Accesibilidad de información sonora»

Se estima que los estudiantes hubieron empleado unas tres horas de formación en accesibilidad multimedia (una de ellas específicamente en subtitulado) antes de realizar la actividad.

De acuerdo al **compromiso ético** del Canal Fundación ONCE en UNED, los datos de los estudiantes se han suministrado anonimizados usando un identificador generado con SHA-512. Además, se han eliminado del estudio los datos de estudiantes que, a pesar de haber realizado la actividad de subtitulado, no dieron su consentimiento para que sus datos fueran utilizados en estudios científicos.

La actividad consistió en ver dos vídeos idénticos de 43 segundos que solo se diferencian en la calidad del subtitulado. El vídeo original fue diseñado de tal forma que su subtitulado presentara características específicas relacionadas con los requisitos incluidos en la norma de subtitulado. Por ejemplo, la existencia de sonidos relevantes para la trama sin correspondencia visual, la existencia de más de un interlocutor, o de diálogos que comprometían la velocidad máxima de subtitulado. Los subtítulos fueron realizados por una experta de FIAPAS (Confederación Española de Familias de Personas Sordas) siguiendo la norma UNE 153010 (ver AENOR 2012). El otro vídeo tenía un subtitulado similar pero se introdujeron pequeñas deficiencias, algunas de ellas inapreciables para alguien que carezca de conocimientos sobre accesibilidad. El orden de los vídeos fue aleatorio, de tal forma que una cohorte (grupo) de alumnos vio primero el vídeo bien subtitulado y luego el mal subtitulado y la otra lo hizo al revés. Después de ver cada uno de los vídeos, los alumnos respondieron a una escala de Likert de 5 niveles y 18 ítems. Los 18 ítems de Likert responden a criterios de la norma UNE 153010 (ver AENOR 2012).

Los términos **escala de Likert** e ítem de Likert se prestan a menudo a confusión ya que se utilizan con distintos significados. En este trabajo se seguirá la convención más habitual (ver Uebersax 2006) de denominar ítem de Likert a cada una de las preguntas de que consta un cuestionario o test, siendo la escala de Likert el conjunto de todos los ítems del cuestionario. Cada ítem se contestó marcando una opción de entre un conjunto ordenado de respuestas o niveles propuesto e idéntico para todos los ítems. Por ello, se debe evitar denominar escala a los niveles de un ítem.

El diseño del experimento fue **triple ciego**. Es decir, a los alumnos no se les informó de si estaban viendo el vídeo con mejor o con peor calidad de subtitulado; los directores del MOOC tampoco conocieron esta información, como tampoco se conocía en el momento de analizar los datos, ya que los vídeos tienen identificaciones ofuscadas con CRC-32b y no contienen ninguna indicación del tipo de subtitulado del vídeo[1]. El «ciego fue liberado» en la fase de elaboración de la discusión de este trabajo (ver Capítulo 6).

[1] En la respuesta a cada ítem, el alumno pudo añadir comentarios. Éstos fueron eliminados en la fase de análisis para que no filtren información referente al tipo de subtitulado que el alumno creyó estar contestando y solo se utilizaron en la fase de discusión (ver Capítulo 6).

3.2 Ficheros suministrados

Se dispuso de los siguientes ficheros `csv`:

- Fichero `grade`: contiene el identificador de estudiante (ofuscado con SHA-512 para no conocer su identidad real) y el grupo al que pertenece (campo `cohort`) ofuscado con CRC-32b.
- Ficheros `test1` y `test2`: son las repuestas a las escalas de Likert sobre la calidad del subtitulado del primer y del segundo vídeo realizado por cada grupo respectivamente.

En la Tabla 3.1 se muestran los 18 ítems de la escala de Likert que se propuso a los alumnos para que evaluaran cada uno de los vídeos. En la Tabla 3.2 se muestran los 5 niveles de cada uno de los ítems de la escala de Likert utilizados para valorar el subtitulado [2].

Tabla 3.1: Ítems de la escala de Likert.

Item	Texto
Q01	La posición de los subtítulos
Q02	El número de líneas por subtítulo
Q03	La disposición del texto respecto a la caja donde se muestran los subtítulos
Q04	El contraste entre los caracteres y el fondo
Q05	La corrección ortográfica y gramatical
Q06	La literalidad
Q07	La identificación de los personajes
Q08	La asignación de líneas a los personajes en los diálogos
Q09	La descripción de efectos sonoros
Q10	La sincronización de las entradas y salidas de los subtítulos
Q11	La velocidad de exposición de los subtítulos
Q12	El máximo número de caracteres por línea
Q13	La legibilidad de la tipografía
Q14	La separación en líneas diferentes de sintagmas nominales, verbales y preposicionales
Q15	La utilización de puntos suspensivos
Q16	La escritura de los números
Q17	Las incorrecciones en el habla
Q18	Los subtítulos del vídeo cumplen en general con los requisitos de accesibilidad

[2] En la codificación original los valores asignados a cada respuesta eran diferentes: la opción «No sé / No contesto» se codificó con 5 y las demás opciones con una unidad menos que la mostrada. En este trabajo se ha hecho una rotación para asignar valores más usuales en la literatura científica sobre el tema.

Tabla 3.2: Niveles de los ítems de la escala de Likert.

values	levels
0	No sé / No contesto
1	Muy en desacuerdo
2	En desacuerdo
3	Neutral
4	De acuerdo
5	Muy de acuerdo

3.3 Preprocesado

Los datos personales de los estudiantes se suministraron anonimizados para evitar conocer su identidad. De acuerdo con el compromiso ético del Canal, del estudio se han eliminado 19 estudiantes que, a pesar de haber realizado la actividad, no dieron su consentimiento para que sus datos se utilizaran en estudios científicos. Tras este proceso, se dispone de 198 cuestionarios correspondientes a 111 alumnos. Hay 24 estudiantes que solo realizaron el primero de los test por lo que se han eliminado del estudio. De estos, 46 manifestaron tener sexo femenino, 19 masculino y el resto (22) prefirieron no suministrar esta información. Se constata que hay un claro sesgo hacia el sexo femenino entre los participantes en la actividad de subtitulado.

En esta sección se describen las transformaciones realizadas con los ficheros suministrados:

- Se leyó el fichero `grade`. El número de fila con el que el estudiante aparece en el fichero se utilizó como identificador del estudiante para mantener la trazabilidad y comprobar que las transformaciones realizadas son correctas.

- Se eliminaron los datos de los estudiantes que, aun habiendo realizado la actividad, no dieron su consentimiento para participar en el estudio.

- El valor del campo `cohort`, que indica el valor anonimizado para el grupo, se sustituyó por una letra, *A* o *B*. En el momento en que se realizó este proceso se desconocía qué vídeo vio primero cada cohorte.

- Se leyeron los ficheros de test y se procesaron. Se utilizó el nombre del fichero (`test1` o `test2`) para saber de qué vídeo se estaba respondiendo el test [3].

[3] Se reitera que en el momento de realizar este proceso se desconocía si el vídeo es el correctamente subtitulado o el otro. La única información que se almacenó es si se estaba respondiendo al vídeo que se vio primero.

- Se seleccionaron los ítems que contienen las respuestas y se renombraron para que fuera más fácil saber de qué ítem se trataba [4]. Se convirtió el campo `LastTry`, que contiene la fecha y hora de realización del test, a formato fecha y hora.

- Se realizaron algunas comprobaciones como la ausencia de valores nulos en las variables más relevantes y la no existencia de inconsistencias o errores de procesado.

- Se eliminaron los comentarios y se grabaron en un fichero aparte para que no revelaran información que habría podido descubrir el tipo de subtitulado que piensa que estaba evaluando el estudiante.

- Se renombraron las variables (ver Tabla 3.3).

- Se eliminaron del estudio los estudiantes que solo han realizado uno de los test.

- Se transformaron las variables que lo requirieron en factores. El ítem 18 se fijó como referencia en el factor `Item` ya que es una valoración general del subtitulado.

- Se rotaron los valores de respuesta para que «No sé / No contesto» tenga valor 0 y el resto de 1 a 5 desde «Muy en desacuerdo», 1, hasta «Muy de acuerdo», 5.

- Se crearon los factores `Level` con los niveles `negativo`, `neutral` y `positivo` dependiendo de si la respuesta es 1 ó 2, 3, 4 ó 5 respectivamente e `Improve` con valores 0 ó 1, dependiendo de si la respuesta en el test *A* es mejor (1) o igual o peor (0) que la del *B* para cada ítem y estudiante.

- Se transformó el `dataframe` de formato ancho a largo. Los ficheros de respuestas originales se suministraron en formato ancho. Es decir, que cada fila es un test que contiene 18 columnas para las respuestas a cada ítem. Los nombres de las columnas son $Q01$, $Q02$, …, $Q18$ y tienen valores de 0 a 5 con las respuestas. La mayoría de los paquetes de R utilizados requieren que los datos estén en formato largo. Esto que quiere decir que cada fila tendrá una única respuesta por lo que habrá únicamente dos columnas, *Item* y *Response*. En la primera se almacena el identificador del ítem ($Q01$, $Q02$, …, $Q18$) y en la segunda el valor de la respuesta (de 0 a 5). De esta forma, un test pasó de ocupar una fila y 18 columnas en el formato ancho a 18 filas y dos columnas en el largo.

[4] En los ficheros suministrados la respuesta a cada ítem ocupaba varios campos. Se seleccionó en cada ítem el que contiene el valor de la respuesta y se convirtió a numérico.

3.4 Variables utilizadas

En la Tabla 3.3 se describen las características más relevantes de las principales variables que se utilizarán en el modelado y en el análisis estadístico. La variable dependiente o respuesta en los modelos ordinales es `Response` y contiene las respuestas a todos los ítems (de 0 a 5). En los modelos logísticos se usa como variable respuesta `Level` o `Improve`. La variable explicativa principal es el factor `Treat` y permite diferenciar los dos niveles de subtitulado (*A* o *B*). Los factores `Period` y `Seq` sirven para evaluar la presencia de efectos periodo y secuencia respectivamente. El factor `Period` toma valores 1 ó 2 en función de si trata del primer o del segundo periodo. El factor `Seq` toma valores *AB* o *BA* dependiendo de si se vio primero el vídeo con subtitulado *A* o con subtitulado *B*. En este trabajo los términos secuencia y grupo se usan indistintamente. Por último, los factores `Subject` e `Item` son variables explicativas que se tratan como efectos aleatorios en el modelado multinivel (ver Sección 4.2) y corresponden respectivamente a los estudiantes y a los ítems de la escala de Likert.

Tabla 3.3: Descripción de las variables más importantes.

Nombre	Descripción	Tipo	Valores
Response	Respuesta a los ítems del test.	Factor ordenado	De 0 a 5
Level	Valoración de la respuesta.	Factor ordenado	Negativo, Neutral, Positivo[1]
Improve	Mejor respuesta en test A que en B.	Factor	1 ó 0
Treat	Subtítulos	Factor	A o B
Period	Periodo	Factor	1 ó 2[2]
Seq	Secuencia de aplicación de los tratamientos.	Factor	AB o BA
Subject	Identificación del estudiante	Factor	Numérico
Item	Número del ítem	Factor	Q01, Q02, ..., Q18

[1] Positivo cuando Response sea 4 ó 5, Negativo cuando sea 1 ó 2 y Neutral para 3.
[2] 1 para el primer vídeo visto y 2 el segundo.

3.5 `Dataframes` utilizados

Se van a usar tres `dataframes` construidos en el preprocesado:

- `df_response` contiene las respuestas con valor de 1 a 5. Se han eliminado las de valor 0 («No sé / No contesto»). Se utiliza cuando se traten las respuestas como ordinales y, por lo tanto, como ordenadas.
- `df_all` incluye todas las respuestas de 0 a 5. Se utiliza cuando se traten las respuestas como categóricas y no ordenadas.
- `df_improve`: Es un `dataframe` en el que la variable respuesta es `Improve` usado en el modelado logístico. Es una variable dicotómica que muestra si la respuesta es mejor en el subtitulado *A* que en el *B* (`Improve`= `Response` *A* > `Response` *B*).

La estructura de los dos primeros dataframes es la siguiente:

```
tibble [2,980 x 7] (S3: tbl_df/tbl/data.frame)
 $ Seq     : Factor w/ 2 levels "AB","BA": 1 1 1 1 1 1 1 1 1 1 ...
 $ Period  : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
 $ Treat   : Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 1 ...
 $ Subject : Factor w/ 87 levels "4","33","35",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Item    : Factor w/ 18 levels "Q18","Q01","Q02",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Response: Ord.factor w/ 5 levels "1"<"2"<"3"<"4"<..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Level   : Ord.factor w/ 3 levels "Negativo"<"Neutral"<..: 2 2 2 2 2 2 2 2 2..
```

En la Tabla 3.4 se muestran algunos ejemplos de datos. Concretamente se muestran las respuestas a tres ítems de dos estudiantes que vieron los vídeos en distinto orden.

Tabla 3.4: Muestra del dataframe preparado para el modelado estadístico en formato largo.

Seq	Period	Treat	Subject	Item	Response	Level
AB	1	A	35	Q18	5	Positivo
AB	2	B	35	Q18	4	Positivo
AB	1	A	35	Q01	5	Positivo
AB	2	B	35	Q01	5	Positivo
AB	1	A	35	Q02	5	Positivo
AB	2	B	35	Q02	4	Positivo
BA	1	B	33	Q18	2	Negativo
BA	2	A	33	Q18	4	Positivo
BA	1	B	33	Q01	4	Positivo
BA	2	A	33	Q01	4	Positivo
BA	1	B	33	Q02	4	Positivo
BA	2	A	33	Q02	4	Positivo

Capítulo 4

Modelado estadístico

Este capítulo comienza con un análisis exploratorio de los datos. Continúa proponiendo diversas formas de adecuar el modelado estadístico que se explicó en el Marco Teórico (ver Capítulo 2) al diseño del experimento del subtitulado. La selección de los modelos se realiza en el Capítulo 5.

4.1 Análisis Exploratorio

Como se explica en la Tabla 3.3, al subtitulado se le denomina tratamiento y a sus niveles (correcto e incorrecto) se les ha llamado *A* y *B* sin hacer ninguna conjetura de cual de los dos es el subtitulado correcto. El grupo con secuencia *AB* será el que primero vio el vídeo con subtitulado *A* y luego el *B*. Análogamente, el grupo con secuencia *BA* vio los vídeos en orden inverso. Recuérdese que el nivel 0 de respuesta se corresponde con «No sé / No contesto» (ver Tabla 3.2). Tras eliminar los test de los 19 estudiantes que no dieron su consentimiento para participar en el estudio y los de los 24 estudiantes que no realizaron el segundo test, las dos cohortes están equilibradas ya que hay 43 estudiantes que realizaron el test con secuencia *AB* y 44 con secuencia *BA*.

Análisis de la calidad de los datos

En esta sección se analiza si hay test que tienen valores de respuesta que puedan resultar anómalos. En los test no se ha observado ningún valor nulo ni erróneo.

El campo `LastTry` contiene la fecha y hora de realización del test. Con esta información se puede conocer el tiempo que transcurrió desde que un estudiante rellenó el primer test hasta que completó el segundo. Dado que los vídeos duran 43 segundos y hay 18 ítems, se puede estimar el tiempo medio que cada estudiante empleó en contestar cada ítem suponiendo que haya visto el segundo vídeo

completo. La Tabla 4.1 muestra que hay algunos test en los que los estudiantes emplearon un tiempo muy breve en contestar [1].

Tabla 4.1: Tiempos de realización de la segunda actividad de duración inferior a 2 minutos.

Subject	Test	Spent time (secs)	Time by Item (secs)
893	B	56	0.72
1020	B	78	1.94
85	A	102	3.28
4	B	103	3.33
110	A	107	3.56
1034	A	118	4.17

La Figura 4.1 muestra que hay 28 test en los que el estudiante contestó a todos los ítems usando únicamente 2 respuestas diferentes. Además hay 13 test en los que se contestaron todos los ítems con 1 respuesta.

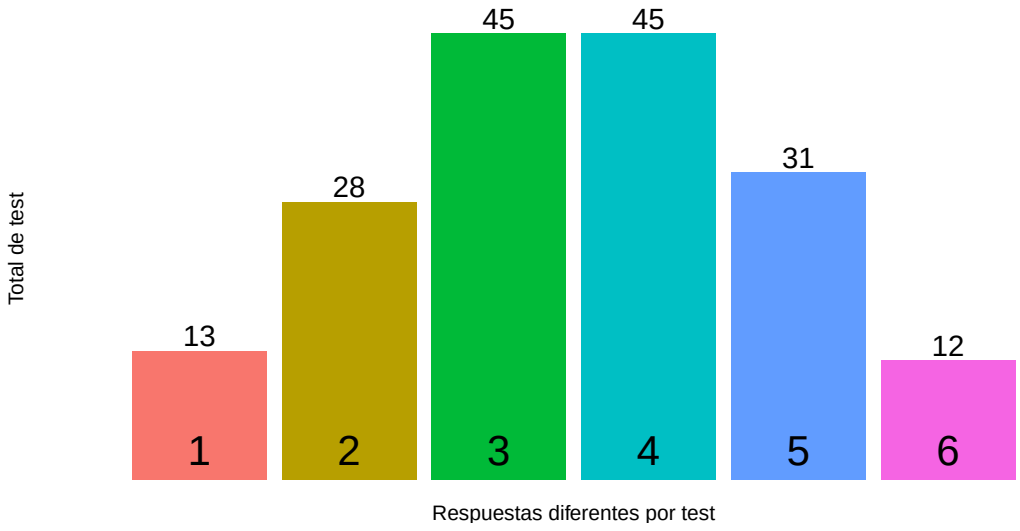

Figura 4.1: Número de respuestas diferentes en un mismo test.

La tabla Tabla 4.2 muestra los test de respuesta única y el valor de esa respuesta. Se aprecia que la mayoría de estos test tienen valor de respuesta 4, la secuencia mayoritaria es la *BA* y el test el *A*. El estudiante 4 responde ambos test utilizando el mismo valor de respuesta en todos los ítems.

[1] Hay que tener en cuenta que la duración de vídeo es de algo más de 40 segundos y que los estudiantes tienen que contestar un test de 18 ítems.

Tabla 4.2: Test en los que todos los ítems se contestan con el mismo valor de respuesta.

Response	Seq	Test	Subject
2	AB	A	4
2	AB	B	4
3	BA	B	734
3	BA	A	803
3	BA	A	33
3	BA	A	229
4	AB	A	35
4	AB	A	76
4	AB	B	523
4	BA	B	85
4	BA	A	901
4	BA	A	808
4	BA	A	871

La Figura 4.2 presenta la distribución de la cantidad de respuestas cuyo valor cambia entre los dos test que realiza cada estudiante. La mayoría de los estudiantes cambian entre uno y otro test entre 11 y 17 respuestas. Tan solo 1 estudiante respondió a todos los ítems con el mismo valor en los dos test. Por otro lado, no hay test que tengan un número excesivo de contestaciones «No sé/No contesto» (ver Tabla 4.3).

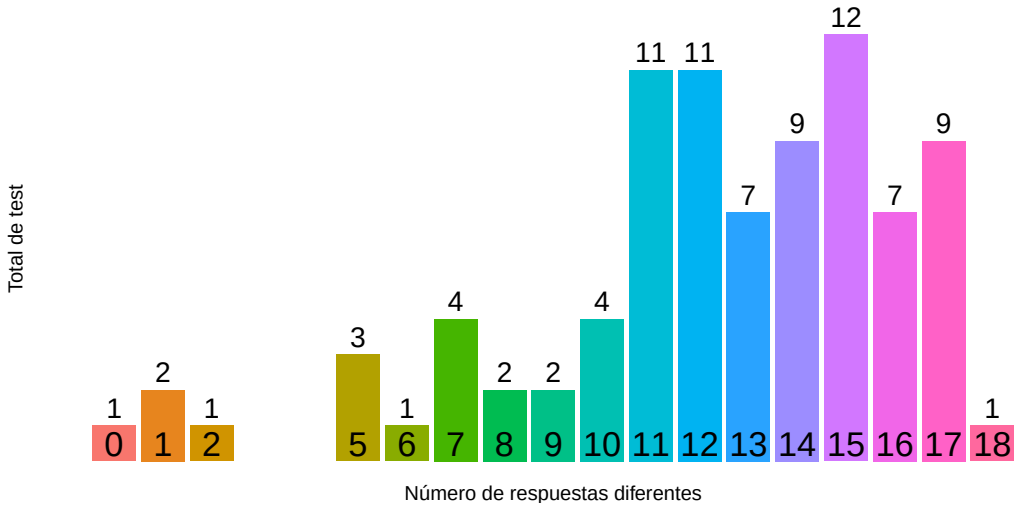

Figura 4.2: Número de respuestas diferentes entre los test para cada estudiante.

Tabla 4.3: Los 5 test con más respuestas 'No sé / No contesto'

Test	Subject	Total answers by test
A	231	5
B	339	5
B	346	5
B	1174	5
A	1187	4

En resumen, se constata que algunos test tienen valores que no parecen muy razonables. Por ejemplo, no parece razonable realizar la actividad en menos de 120 segundos. Además en algunos test hay poca variabilidad en las respuestas. Sin embargo, no son muchos los test con estas características así que se ha decidido mantener estos datos a pesar de que se pueda dudar de si en ellos los estudiantes contestaron con la debida atención y diligencia.

Comparación de los subtítulos *A* y *B* entre grupos

La Figura 4.3 presenta una forma de comparar los dos test realizados por los estudiantes. Para cada estudiante se comparó ítem a ítem sus dos test y se contabilizó la diferencia entre el número de ítems en los que la puntuación en el segundo vídeo fue superior y en los que lo fue inferior (las que no variaron de puntuación no se consideraron). En el eje x se muestran las diferencias entre respuestas. Cantidades negativas indican que hay más respuestas en el segundo de los test que han empeorado respecto al primero de las que han mejorado. En el eje y se representa el número de estudiantes para cada diferencia. Esta frecuencia se representa en negativo cuando la diferencia en el eje x sea negativa para facilitar la comparación[2]. Esto es una forma de evaluar si el estudiante valoró mejor el segundo vídeo que el primero.

Se aprecia que en el grupo *AB* las diferencias tienden a ser negativas y en el *BA* positivas. Esto estaría indicando que los estudiantes valoran mejor el subtitulado de nivel *A* en ambas secuencias. Por ello, es esperable que las respuestas de los estudiantes del grupo *AB* hayan empeorado y que las diferencias sean negativas y que lo contrario haya sucedido con las del grupo *BA*. La diferencia más frecuente en el grupo *AB* es 12 y en el grupo *BA* este valor es 11. Resulta llamativo que haya estudiantes cuyas contestaciones estén tan alejadas de la tendencia de su grupo. En la Tabla 4.4 se muestran los tiempos que han transcurrido entre la realización de los test de aquellos estudiantes cuyas respuestas difieren de forma importante de su grupo. Son aquellos que aparecen en azul en la secuencia *AB* y en rojo en la secuencia *BA*. Se observa que casi todos son tiempos entre actividades muy cortos. En cualquier caso y, como no son muchos, se ha decidido no eliminarlos y realizar el análisis con ellos.

[2]En la comparación se han omitido aquellas respuestas en las que el estudiante contestó «No sé / No contesto» en el ítem correspondiente de uno de los test.

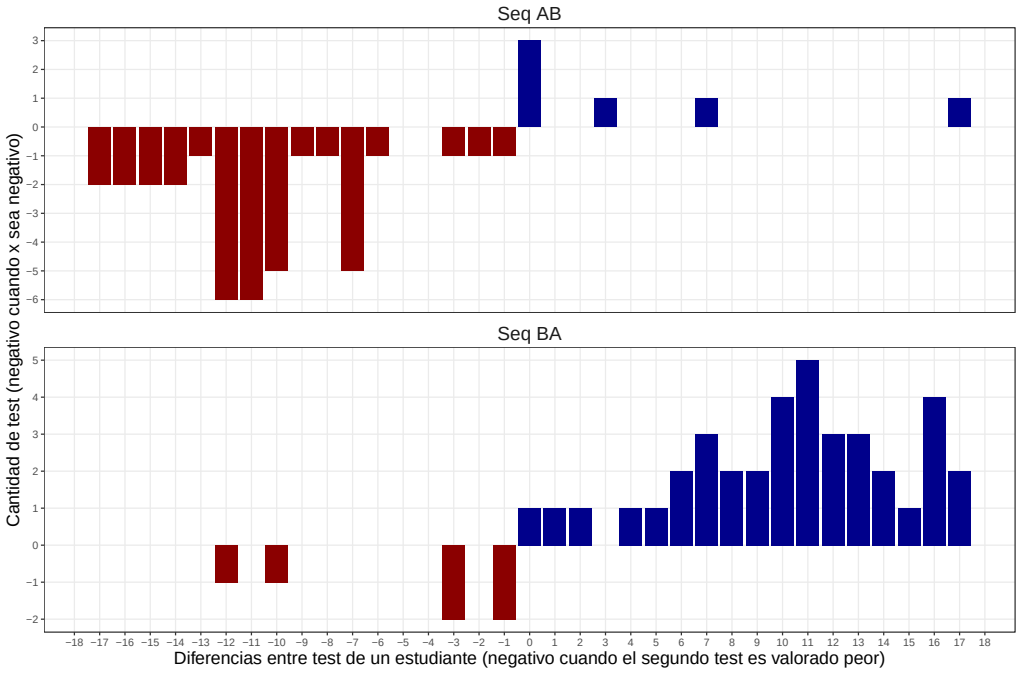

Figura 4.3: Diferencias en las respuestas entre test por estudiante y grupo.

Tabla 4.4: Estudiantes que tienen diferencias en sus respuestas muy alejadas de la tendencia de su grupo.

Seq	Subject	Diff	Minutes
AB	1020	17	1.3
AB	75	7	3.33
BA	650	-10	50345.95
BA	85	-12	1.7

Tabla 4.5: Resumen de frecuencias de respuesta.

				Response				
Seq	Period	Treat	0	1	2	3	4	5
AB	1	A	39	2	25	71	203	434
AB	2	B	43	87	185	121	172	166
BA	1	B	40	76	174	127	237	138
BA	2	A	30	2	30	64	345	321

En la Tabla 4.5 se muestra la frecuencia absoluta del valor de respuesta para cada grupo y test en todos los ítems. Esta es otra forma de comparar los niveles de subtitulado. La Figura 4.4 muestra la misma información gráficamente y con frecuencias relativas. En esta figura se pueden apreciar algunas cuestiones interesantes:

- El tratamiento (subtitulado) con nivel *A* presenta claramente mayores valores de respuesta que el *B* como ya se había visto (ver Figura 4.3).

- En general los dos grupos (*AB* y *BA*) muestran bastante acuerdo en el subtitulado en ambos niveles: En el nivel de tratamiento *A* los dos grupos tienen una frecuencia relativa similar de respuestas positivas (valores 4 y 5). El grupo *AB* tiene un 82% de respuestas positivas y el grupo *BA* 84%. No obstante, el grupo *AB* tiene más respuestas con valor 5 que el grupo *BA* (56% frente a 41%). La valoración es también similar entre grupos en el nivel de tratamiento *B*: el grupo *AB* tiene 44% de respuestas positivas y 47% el grupo *BA*. Las valoraciones negativas (1, 2), la neutra (3) y la "No sé / No contesto" (0) son también muy similares en ambos grupos.

Figura 4.4: Frecuencias relativas de las respuestas al test.

El análisis marginalizado de tratamiento, secuencia y periodo tiene estos resultados referidos a los ítems con contestación positiva (4, 5):

- El tratamiento *A* tiene un 83% marginalizado de respuestas positivas frente al 46% del tratamiento *B*.
- El periodo 1 tiene un 65% marginalizado de respuestas positivas frente al 64% del periodo 2.
- Finalmente, la secuencia *AB* tiene un 63% de respuestas positivas frente 66% de la secuencia *BA*.

Análisis de los ítems

El gráfico Figura 4.5 muestra la frecuencia relativa por grupo y por test de los ítems clasificados por niveles de respuesta, considerando que:

- Los niveles 1 y 2 se consideran valoraciones negativas.
- El nivel 3 se considera neutro.
- Los niveles 4 y 5 se consideran positivos.
- El nivel 0 («No sé / No contesto») se excluye en este análisis.

Se muestra en primer lugar el ítem 18 por ser una valoración global del subtitulado y que resume la opinión que sobre el mismo tiene el estudiante. Se vuelve a constatar que el subtitulado *A* es mejor valorado por los estudiantes, pero ahora se confirma que en los 18 ítems ambos grupos tienen más puntuaciones positivas y menos negativas en el subtitulado *A* que en el *B*. También se vuelve a constatar que los dos grupos valoran de forma muy similar los dos niveles de subtitulado en todos los ítems. En el nivel de subtitulado *A* los ítems *Q*15, *Q*16 y *Q*17 obtienen relativamente peores valoraciones (consultar la Tabla 3.1 para ver el texto de los ítems) y estas son similares en ambos subtitulados. Hay algunos ítems que son valorados de forma muy positiva incluso en el nivel de subtitulado *B* (por ejemplo *Q*04 o *Q*13). Por último, los ítems *Q*05 y *Q*09 (también la *Q*14 pero solo para el grupo *BA*) tienen una valoración muy negativa en el nivel de subtitulado *B*.

En la Tabla 4.6 se muestran las contestaciones «No sé / No contesto» por subtitulado e ítem [3]. Se observa que hay relativamente pocas contestaciones «No sé / No contesto» y que éstas se concentran en los ítems *Q*14, *Q*15, *Q*16 y *Q*17. El número de respuestas «No sé / No contesto» está razonablemente equilibrado entre subtitulados excepto en el ítem *Q*10.

Tabla 4.6: Contestaciones “No sé / No contesto” por nivel de subtitulado e ítem

Q03	Q05	Q06	Q08	Q09	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17
A												
1	0	1	3	1	0	1	1	0	11	11	22	17
B												
4	1	0	1	0	8	2	4	1	10	15	25	12

4.2 Modelos utilizados

Es esta sección se concreta la forma de aplicar los modelos presentados en el Marco teórico (ver Capítulo 2) en la actividad de subtitulado.

[3] Solo se muestran los ítems que tienen respuestas «No sé / No contesto» en alguno de los subtitulados.

Figura 4.5: Frecuencias relativas de las respuestas por ítem.

Comparación con *Odds Ratio* [4]

La métrica *odds ratio* (OR) permite medir la asociación entre dos variables con dos niveles cada una. En el diseño de experimento que se está analizando, los factores $Treat$, $Period$ y Seq tienen todos 2 niveles y se puede contrastar si hay interacción entre cada par de factores para cada nivel de respuesta. Es decir, se contrasta la hipótesis $H_0 : OR = 1$ de ausencia de asociación frente a $H_1 : OR \neq 1$ de existencia de asociación en algún nivel de respuesta. Por ejemplo, el OR para el nivel respuesta r entre subtítulos y secuencias se define de la siguiente forma:

$$OR_{(Treat,Seq|Response=r)} = \frac{\frac{P(Treat=A|Seq=AB,Response=r)}{P(Treat=B|Seq=AB,Response=r)}}{\frac{P(Treat=A|Seq=BA,Response=r)}{P(Treat=B|Seq=BA,Response=r)}}$$

Si los *odds* son similares en cada nivel de respuesta, se acepta la hipótesis nula de que los grupos responden de forma similar a cada nivel de subtitulado y secuencia. En la Sección 5 se pueden consultar los resultados obtenidos. En esta misma sección se hace un test similar pero entre subtitulado y periodos. Para

[4] Esta técnica se ha omitido en el Marco teórico por considerarla conocida por el lector. Si se desea ampliar información se puede consultar Agresti (2010, p. 18).

realizar el contraste de hipótesis se usa la función *loddsratio* del paquete `vcd` (ver Zeileis et al. 2007).

Regresión Logística

En la Sección 2.2 se presentó el fundamento teórico de la Regresión Logística. En esta sección se justifica el uso de este modelo y se ajustan y comparan varios modelos. La variable respuesta se compone de 5 valores ordenados. Esto imposibilita usar directamente la Regresión Logística ya que requiere que la variable de respuesta sea dicotómica. No obstante, se puede comparar la respuesta que cada estudiante dio a cada uno de los subtitulados y comprobar si ha mejorado. Esto producirá una variable de respuesta binaria que permitirá el uso de la Regresión Logística. No obstante, esta transformación reducirá la cantidad de datos disponibles a la mitad e impedirá analizar el efecto periodo ya que al comparar los subtitulados, desaparece el periodo. Se ha creado una variable `Improve` con dos valores posibles: 1 cuando el estudiante valoró el ítem mejor en el subtitulado *A* que en el *B*, 0 si empeoró o puntuó igual. Si en uno de los test contestó un ítem con «No sé / No contesto», se elimina ese ítem.

Se ajusta el modelo con la secuencia como predictor:

```
glm_improve_seq <- glm(Improve ~ 1 + Seq, family = "binomial", data = df_improve)
summary(glm_improve_seq)
```

```
Call:
glm(formula = Improve ~ 1 + Seq, family = "binomial", data = df_improve)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.37858    0.07635   4.958 7.11e-07 ***
SeqBA       -0.05137    0.10668  -0.481     0.63
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1967.2  on 1450  degrees of freedom
Residual deviance: 1966.9  on 1449  degrees of freedom
AIC: 1970.9

Number of Fisher Scoring iterations: 4
```

Se constata que el coeficiente del intercepto es positivo y significativo (0.38). El intercepto es el *log odds* de mejorar la valoración en *A* sobre *B* respecto a empeorar la valoración. La probabilidad de que la respuesta a un ítem sea mejor en el subtitulado *A* que en el *B* es 0.59. La secuencia no resulta significativa y además añadirla apenas reduce la «deviance», por lo que el modelo nulo sin predictores resulta más parsimonioso.

Otra forma de plantear una Regresión Logística es crear una variable de respuesta dicotómica que tenga valor 1 cuando la respuesta sea positiva (valores 4 ó 5)

y cero cuando no lo sea (valores 1, 2 ó 3). En la Sección 5 se comentan los resultados de este modelo.

Regresión Ordinal

En la Sección 2.2 se presentó el fundamento teórico de la Regresión Ordinal Acumulativa (*CM*). En esta sección se comprueban las hipótesis de este modelo para el experimento del subtitulado de vídeos y se ajustan varios modelos que tratan de predecir el nivel de respuesta (variable `Response`) obtenido en cada uno de los ítems de Likert. Concretamente, se compara el modelo que tenga como único predictor el nivel de subtitulado (`Treat`) con el modelo nulo (sin predictores) y también con el modelo en el que se han añadido los predictores `Period` y `Seq` para comprobar si hay significación estadística de la presencia de efectos periodo y secuencia respectivamente.

Comprobación de las hipótesis del modelo *CM*

El modelo *CM* presupone que los *odds* entre dos niveles de respuesta son proporcionales para los mismos valores de variables explicativas. Como se vio en la Ecuación 2.4, es equivalente comprobar que los *odds* son proporcionales que comprobar que la diferencia en `logits` es constante.

No existe un acuerdo generalmente aceptado sobre como comprobar la proporcionalidad de *odds*. Así, por ejemplo, el paquete `ordinal` (ver Christensen 2022) dispone de la función `nominal_test()` que lo que hace es realizar un test de razón de verosimilitud para cada predictor ajustando un modelo en el que se ha relajado la condición de proporcionalidad. Otra posibilidad es utilizar el Test de Brant (ver Brant 1990) que compara los coeficientes obtenidos con los que resultarían de ajustar cada nivel de respuesta mediante una Regresión Logística. Finalmente Harrell (2015, ver pp. 315-316) propone un método gráfico para verificar la hipótesis de proporcionalidad de *odds*. En este trabajo se ha preferido esta última técnica. Para ello se calcula la diferencia en `logits` acumulados entre dos niveles de respuesta consecutivos en cada valor de cada variable predictiva y se comprueba si las diferencias son similares. En la Figura 4.6 se han calculado para los predictores `Treat`, `Period`, `Seq` y `Item` las diferencias de `logits` entre cada dos niveles consecutivos de respuesta. Se constata que las diferencias son pequeñas particularmente para el periodo y para la secuencia. También son moderadas para la mayoría de los ítems. La diferencia es mayor en el subtitulado en la comparación de los niveles de respuesta 1 y 2. Con esta evidencia, se acepta la hipótesis de proporcionalidad de *odds*. En la Tabla 4.7 se muestra como se realiza el cálculo de la diferencia de *odds* para el predictor `Seq` y así facilitar la comprensión de la construcción de la figura. En caso de que la proporcinalidad de *odds* no se cumpla existen varias posibilidades. Una sería desechar el modelo ordinal y usar una Regresión Multinomial. Otra sería relajar la hipótesis de proporcionalidad de *odds* estimando un coeficiente distinto para cada nivel de respuesta y nivel de factor. La función `vglm` del paquete `VGAM` (ver Yee 2023) permite hacer esto.

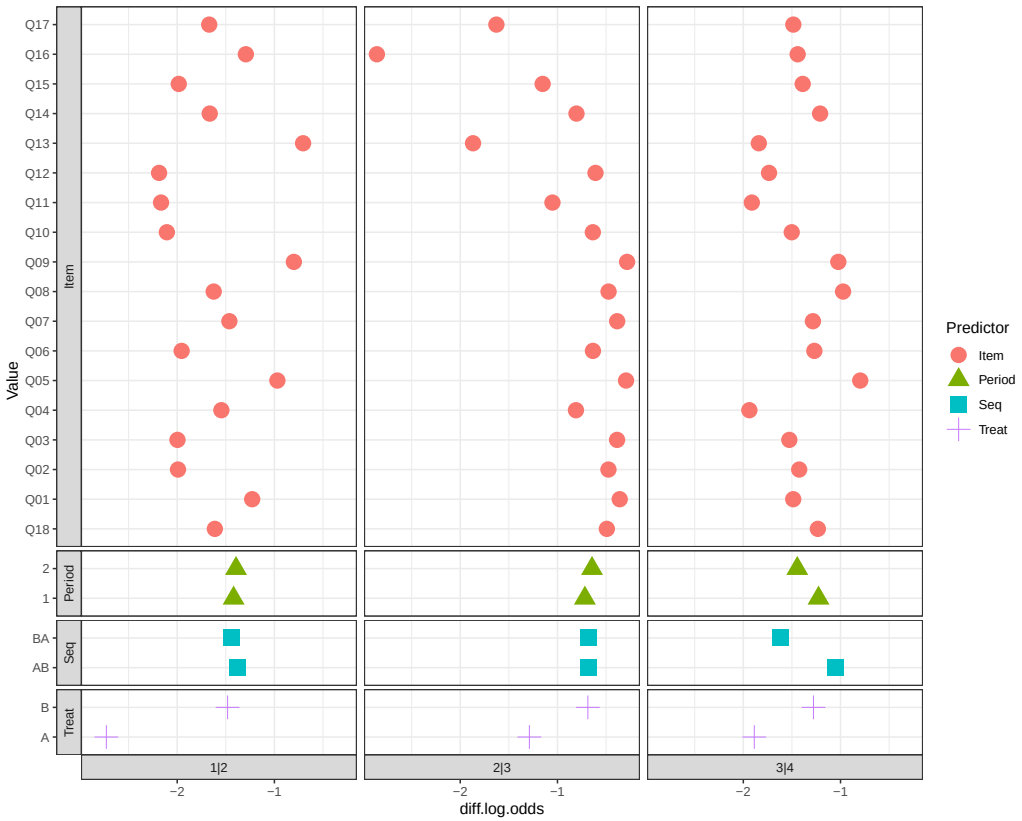

Figura 4.6: Comprobación de la proporcionalidad de *odds*.

Tabla 4.7: Comprobación de la proporcionalidad de *odds* para Seq.

Response	n	cum.sum	odds	log.odds	diff.log.odds
AB					
1	89	89	0.06	−2.74	−1.38
2	210	299	0.26	−1.36	−0.68
3	192	491	0.50	−0.69	−1.05
4	375	866	1.44	0.37	*−Inf*
5	600	1466	*Inf*	*Inf*	NA
BA					
1	78	78	0.05	−2.91	−1.44
2	204	282	0.23	−1.47	−0.69
3	191	473	0.45	−0.79	−1.62
4	582	1055	2.30	0.83	*−Inf*
5	459	1514	*Inf*	*Inf*	NA

Ajuste del modelo ordinal `Response ~ Treat`

Existen varios paquetes en R que permiten ajustar un modelo *CM* con función de enlace logística. El más popular es el paquete `ordinal` (ver Christensen 2022).

El paquete `VGAM` (ver Yee 2023) es más flexible y potente. Otra posibilidad es usar la función `polr` del paquete `MASS` (ver Venables y Ripley 2002). Finalmente la función `orm` del paquete `rms` también permite hacerlo (ver Harrell 2015). En este trabajo se usa el paquete `ordinal` (ver Christensen 2022) por permitir también incluir efectos aleatorios que se utilizarán en el modelado multinivel. Se comienza con un modelo simple que tiene como único predictor el nivel de subtitulado por ser la variable más importante al ser el objeto de la pregunta de investigación:

$$\text{logit}(P(Response_i \leq k)) = \tau_k - \beta_1 \text{Treat}_i,$$

```
clm_treat <-
    clm(
        Response ~ Treat,
        data = df_response, link = "logit"
    )
summary(clm_treat)
```

```
formula: Response ~ Treat
data:    df_response

 link  threshold nobs logLik   AIC     niter max.grad cond.H
 logit flexible  2980 -3966.11 7942.21 5(0)  1.64e-10 3.1e+01

Coefficients:
       Estimate Std. Error z value Pr(>|z|)
TreatB  -1.7206     0.0731  -23.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
    Estimate Std. Error z value
1|2 -3.97230    0.09678 -41.045
2|3 -2.45446    0.06812 -36.029
3|4 -1.66453    0.05936 -28.042
4|5 -0.10547    0.04946  -2.132
```

La función `summary()` muestra la información resumen. Para su interpretación se va a seguir Christensen (2018). El número de condición Hessiano es inferior a 10^4 lo que es indicativo de que no hay problemas de optimización [5]. La sección de coeficientes es la más importante: Se muestra la estimación de parámetros, el error estándar y la significación estadística de acuerdo al Test de Wald para el parámetro `TreatB`. Se comprueba que el valor es claramente significativo. Es decir, que los estudiantes han valorado de forma diferente la calidad del subtitulado en ambos vídeos. El estimador de maxima verosimilitud del coeficiente `TreatB` es -1.72. Siguiendo la deducción de Bruin (2011) se puede, por ejemplo, hacer la siguiente interpretación del significado de este coeficiente referido a dos niveles consecutivos de respuesta, por ejemplo 1 y 2:

[5] El número de condición de Hessiano es una medida de la curvatura de una función en un punto. Si el número de condición de Hessiano es grande, la función es muy sensible a pequeñas perturbaciones y puede ser difícil de optimizar.

$$\begin{aligned} logit[P(Y \leq 1)] &= -3.97 - (-1.72x_1) \\ logit[P(Y \leq 2)] &= -2.45 - (-1.72x_1) \end{aligned}$$

Por lo tanto y teniendo en cuenta que $x_1 = 1$ cuando $Treat = B$ y $x_1 = 0$ cuando $Treat = A$, se pueden calcular los *odds* de A y de B:

$$\begin{aligned} \frac{P(Y \leq 1 \mid x_1 = B)}{P(Y > 1 \mid x_1 = B)} &= exp(-3.97)/exp(-1.72) \\ \frac{P(Y \leq 1 \mid x_1 = A)}{P(Y > 1 \mid x_1 = A)} &= exp(-3.97) \\ \frac{P(Y \leq 2 \mid x_1 = B)}{P(Y > 2 \mid x_1 = B)} &= exp(-2.45)/exp(-1.72) \\ \frac{P(Y \leq 2 \mid x_1 = A)}{P(Y > 2 \mid x_1 = A)} &= exp(-2.45) \end{aligned}$$

Y los *OR* del subtitulado B sobre A para los niveles de respuesta 1 y 2:

$$\begin{aligned} \frac{P(Y \leq 1|x_1 = B)}{P(Y > 1|x_1 = B)} \Big/ \frac{P(Y \leq 1|x_1 = A)}{P(Y > 1|x_1 = A)} &= 1/exp(-1.72) = 5.59 \\ \frac{P(Y \leq 2|x_1 = B)}{P(Y > 2|x_1 = B)} \Big/ \frac{P(Y \leq 2|x_1 = A)}{P(Y > 2|x_1 = A)} &= 1/exp(-1.72) = 5.59 \end{aligned}$$

Se comprueba que el *OR* es equivalente en todos los niveles de respuesta al cuestionario. Esta es la suposición principal de los modelos *CM*. El *odds* de respuesta al cuestionario entre los niveles inferiores y superiores a uno dado, k, es 5.59 veces en el subtitulado B que en el A. Esto indica que el subtitulado B es percibido por los estudiantes como de peor calidad que el subtitulado A. Concretamente, el *OR* de observar una mejor respuesta en un ítem del test es 5.59 veces superior en el nivel de subtitulado A que en el B. Aunque no suele ser de interés, la interpretación de los coeficientes de los umbrales (`Threshold coefficients`), se pueden utilizar para estimar las probabilidades de respuesta. Por ejemplo, para el nivel de subtitulado B y nivel de respuesta 2:

$$\begin{aligned} logit[P(Y \leq 1)] &= -3.97 - (-1.72) = -2.25 \\ P(Y \leq 1) &= \frac{exp(-2.25)}{1 + exp(-2.25)} = 0.10 \\ logit[P(Y \leq 2)] &= -2.45 - (-1.72) = -0.73 \\ P(Y \leq 2) &= \frac{exp(-0.73)}{1 + exp(-0.73)} = 0.32 \\ P(Y = 2) &= P(Y \leq 2) - P(Y \leq 1) = 0.23 \end{aligned}$$

Para el subtitulado A no se tiene en cuenta el coeficiente $TreatB$ ya que el valor x_1 es cero:

Tabla 4.8: Probabilidades de respuesta para el modelo ordinal Response ~ Treat

	1	2	3	4	5
A	0.018	0.061	0.08	0.315	0.526
B	0.095	0.229	0.19	0.320	0.166

$$
\begin{aligned}
logit[P(Y \leq 1)] &= && -3.97 \\
P(Y \leq 1) &= \frac{exp(-3.97)}{1 + exp(-3.97)} = & 0.02 \\
logit[P(Y \leq 2)] &= && -2.45 \\
P(Y \leq 2) &= \frac{exp(-2.45)}{1 + exp(-2.45)} = & 0.08 \\
P(Y = 2) &= P(Y \leq 2) - P(Y \leq 1) = & 0.06
\end{aligned}
$$

En Tabla 4.8 se muestran las probabilidades para ambos niveles de subtitulado y todos los posibles valores de respuesta. Se confirma que en el nivel de subtitulado *A* son más probables las respuestas 5 y 4, siendo poco probables el resto de niveles. Sin embargo, en el subtitulado *B* existe bastante incertidumbre, siendo el valor más probable el nivel 4 y muy similares los niveles 2, 3 y 5. Esto se corresponde con lo que ya se había constatado en el Análisis Exploratorio (ver Figura 4.4). Se debe tener en cuenta que este modelo tiene un único predictor y, por lo tanto, no es capaz de explicar las diferencias en el nivel de respuesta para distintos periodos, secuencias, ítems o estudiantes. En las siguientes secciones se investiga si en el nivel de respuesta influyen estos predictores.

Ajuste del modelo ordinal `Response ~ Treat * Period`

Para saber si existe un efecto periodo, se añade como predictor la variable `Period`. También se añade la interacción entre subtitulado y periodo [6]:

$$\text{logit}(P(Response_i \leq k)) = \tau_k - \beta_1 \text{Treat}_i - \beta_2 \text{Period}_i - \beta_3 \text{Treat}_i : \text{Period}_i \quad (4.1)$$

En el Apéndice A se demuestra que cuando el contraste es *sum* la interacción entre periodo y subtitulado es equivalente al efecto secuencia. Es decir, que los modelos `Response ~ Treat*Period` y `Response ~ Treat + Period + Seq` son equivalentes. Esto no sucede cuando el contraste es *treatment*, que es el utilizado por defecto en R. En la Tabla 4.9 se comparan los coeficientes de los cuatro modelos que se listan a continuación:

- `Response ~ Treat * Period` con contraste `treatment`.

[6] Se debe tener en cuenta que en R la interacción entre dos variables se puede añadir con los símbolos «*» y «:». El símbolo «*» añade al modelo tanto los efectos principales como la interacción, mientras que el símbolo «:» tan solo añade la interacción. Por ello, los modelos $Response \sim Treat * Period$ y $Response \sim Treat + Period + Treat : Period$ son equivalentes en R

- `Response ~ Treat + Period + Seq` con contraste `treatment`.
- `Response ~ Treat * Period` con contraste `sum`.
- `Response ~ Treat + Period + Seq` con contraste `sum`.

Tabla 4.9: Comparación de los coeficientes con contraste "treatment" y "sum".

contr.treatment				contr.sum			
Response ~ Treat*Period		Response ~ Treat+Period+Seq		Response ~ Treat*Period		Response ~ Treat+Period+Seq	
coef	value	coef	value	coef	value	coef	value
1|2	−4.246	1|2	−4.246	1|2	−3.127	1|2	−3.127
2|3	−2.728	2|3	−2.728	2|3	−1.608	2|3	−1.608
3|4	−1.938	3|4	−1.938	3|4	−0.818	3|4	−0.818
4|5	−0.370	4|5	−0.370	4|5	0.750	4|5	0.750
TreatB	−1.960	TreatB	−1.748	Treat1	0.874	Treat1	0.874
Period2	−0.492	Period2	−0.279	Period1	0.140	Period1	0.140
TreatB:Period2	0.425	SeqBA	−0.213	Treat1:Period1	0.106	Seq1	0.106

Se comprueba que coinciden los coeficientes de los dos modelos con contraste `sum` y que el efecto secuencia es equivalente a la interacción de periodo y subtitulado con este contraste. Sin embargo, en el contraste `treatment` coinciden los coeficientes de los interceptores pero no así los de los factores. Además, estos tres últimos coeficientes tienen nombres diferentes en los dos contrastes. La diferencia en el nombre se corresponde con la distinta interpretación del significado de los coeficientes. En el contraste `treatment` los valores de los interceptos se refieren a los valores de los factores en el nivel de referencia de cada factor (en este caso $Treat = A$ y $Period = 1$) y los valores de los otros coeficientes ($TreatB$ y $Period2$) son la diferencia con el de referencia. Así, por ejemplo, $TreatB$ es la diferencia con $TreatA$ en el periodo 1. Con este tipo de contraste es más difícil aislar el efecto que produce un nivel de un factor independiente del otro factor. En el contraste `sum` los valores de los interceptos son el efecto medio [7], y los coeficientes $Treat1$ y $Period1$ son los efectos que sobre ese valor medio produce el nivel de factor de referencia, que en este caso es el primero ($Treat = A$ y $Period = 1$ respectivamente). Así por ejemplo en el contraste `sum`:

- El coeficiente 1|2 tiene un valor -3.127 y es el `logit` medio de que la respuesta sea menor que 1 frente a que sea mayor que 1.
- El coeficiente $Treat1$ tiene un valor de 0.874 y es la diferencia en `logits` que se añade en el nivel de subtitulado A sin tener en cuenta el periodo. Es decir, que es el efecto del subtitulado A. Su valor es positivo. Como en la Ecuación 4.1 aparece restando, el subtitulado A hace más pequeño el `logit` y, por lo tanto, disminuye la probabilidad de una respuesta inferior frente a una superior.
- Para obtener el efecto del subtitulado B se cambia el signo a $Treat1$: -0.874. Por ello aumenta la probabilidad de un menor valor de respuesta.
- La diferencia en `logits` de los efectos totales del subtitulado es el doble de 0.874.

[7] Se calcula como la media de las medias de cada combinación de los niveles de factor.

- El coeficiente $Period1$ tiene un valor 0.14 y es la diferencia en `logits` que produce el periodo 1 sin tener en cuenta el subtitulado.
- El efecto del periodo 2 se obtiene cambiado el signo al efecto del periodo 1: -0.14.
- El efecto total del periodo es 0.279 `logits`.
- El coeficiente $Treat1 : Period1$ tiene un valor de 0.106 y es la interacción entre el subtitulado A y el periodo 1. Es equivalente al efecto en `logits` de la secuencia AB. El efecto de la secuencia BA será -0.106.
- Por lo tanto el efecto total en `logits` del subtitulado A en el periodo 1 será $1|2 - Treat1 - Period1 - Treat1 : Period1 = -3.127 - 0.874 - 0.14 - 0.106 = -4.246$. Obsérvese que este valor corresponde con el parámetro 1|2 de los modelos con contraste `treatment`.
- El efecto total en `logits` del subtitulado B en el periodo 1 será $1|2 + Treat1 - Period1 + Treat1 : Period1 = -3.127 + 0.874 - 0.14 + 0.106$.
- El efecto total en `logits` del subtitulado A en el periodo 2 será $1|2 - Treat1 + Period1 + Treat1 : Period1 = -3.127 - 0.874 + 0.14 + 0.106$.
- El efecto total en `logits` del subtitulado B en el periodo 2 será $1|2 + Treat1 + Period1 - Treat1 : Period1 = -3.127 + 0.874 + 0.14 - 0.106$.

En la Tabla 4.10 se muestra la equivalencia de los coeficientes entre los modelos ajustados con cada contraste. La conclusión que se obtiene de todo esto es que cuando se usan dos o más factores, la interpretación con contraste `sum` resulta más intuitiva y sencilla y será el contraste utilizado en este trabajo.

Tabla 4.10: Equivalencia entre los coeficientes `contr.treatment` y `contr.sum` en el modelo Response ~ Treat*Period.

contr.treatment	contr.sum	value
1|2	1|2 - Treat1 - Period1 - Treat1:Period1	-4.246
2|3	2|3 - Treat1 - Period1 - Treat1:Period1	-2.728
3|4	3|4 - Treat1 - Period1 - Treat1:Period1	-1.938
4|5	4|5 - Treat1 - Period1 - Treat1:Period1	-0.37
TreatB	-2(Treat1 + Treat1:Period1)	-1.96
Period2	-2(Period1 + Treat1:Period1)	-0.492
TreatB:Period2	4(Treat1:Period1)	0.425

A continuación se muestra el resumen del modelo con contraste `sum` para constatar que los tres coeficientes son significativos:

```
formula: Response ~ Treat * Period
data:    df_response

 link  threshold nobs logLik   AIC     niter max.grad cond.H
 logit flexible  2980 -3953.01 7920.03 5(0)  2.13e-10 1.4e+01

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
Treat1        0.87395    0.03678  23.763  < 2e-16 ***
Period1       0.13962    0.03411   4.094 4.25e-05 ***
```

```
Treat1:Period1  0.10627    0.03410   3.117  0.00183 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
    Estimate Std. Error z value
1|2 -3.12665    0.08242  -37.94
2|3 -1.60838    0.04968  -32.38
3|4 -0.81849    0.04225  -19.37
4|5  0.74993    0.04194   17.88
```

Elección del modelo ordinal mediante el test de razón de verosimilitud

La Tabla 4.11 compara tres modelos ordinales con el contraste *sum*:

- Modelo nulo.
- Modelo con predictor `Treat`.
- Modelo con predictores `Treat` y `Period` y su interacción (que es equivalente a incluir el predictor `Seq`).

Se constata que los coeficientes estimados en los tres modelos son significativos y de similar valor.

Tabla 4.11: Comparación de modelos ordinales.

	Response ~ 1		Response ~ Treat		Response ~ Treat:Period	
	Est.	S.E.	Est.	S.E.	Est.	S.E.
1|2	-2.824***	0.080	-3.112***	0.082	-3.127***	0.082
2|3	-1.418***	0.046	-1.594***	0.049	-1.608***	0.050
3|4	-0.738***	0.039	-0.804***	0.042	-0.818***	0.042
4|5	0.596***	0.038	0.755***	0.042	0.750***	0.042
Treat1			0.860***	0.037	0.874***	0.037
Period1					0.140***	0.034
Treat1 × Period1					0.106**	0.034
Num.Obs.	2980		2980		2980	
AIC	8541.7		7942.2		7920.0	
BIC	8565.7		7972.2		7962.0	
Log.Lik.	-4266.851		-3966.107		-3953.013	
edf	4		5		7	

Al ser los tres modelos anidados se pueden comparar con la prueba de razón de verosimilitud. Se comprueba que el tercer modelo reduce significativamente el logaritmo de la función de verosimilitud y, por lo tanto, debe ser aceptado:

```
Likelihood ratio tests of cumulative link models:

                     formula:                  link: threshold:
clm_sum_null         Response ~ 1              logit flexible
clm_sum_treat        Response ~ Treat          logit flexible
clm_sum_treat.period Response ~ Treat * Period logit flexible

                     no.par    AIC  logLik LR.stat df Pr(>Chisq)
clm_sum_null              4 8541.7 -4266.9
clm_sum_treat             5 7942.2 -3966.1 601.490  1  < 2.2e-16 ***
clm_sum_treat.period      7 7920.0 -3953.0  26.186  2  2.059e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Este modelo estima coeficientes positivos para `Treat1`, `Period1` y `Treat1:Period1` (equivalente a `Seq1`). Estos coeficientes indican que:

- Son más probables mayores niveles de respuesta en el subtitulado *A* que en el *B*.
- Son más probables mayores niveles de respuesta en el periodo 1 que en el 2.
- Son más probables mayores niveles de respuesta en la secuencia *AB* que en la secuencia *BA*.
- No obstante, y a pesar de que el efecto periodo y el efecto secuencia son significativos, el efecto del nivel de subtitulado medido en `logits` es ocho veces más importante que estos efectos considerados individualmente y cuatro veces considerados de forma conjunta.

Regresión Ordinal Multinivel

En la Sección 2.3 se expuso el fundamento teórico de los modelos multinivel. Aquí se justifica su interés aplicado al caso del subtitulado de vídeos. Hay dos variables susceptibles de ser incorporadas al modelo como efectos aleatorios. El primer candidato es el factor `Subject`. Es evidente que los estudiantes son una muestra de una población más amplia que estaría constituida por todos los estudiantes del curso de accesibilidad. Pero es que además cada estudiante responde a cada ítem dos veces y, por lo tanto, sus observaciones no son independientes. En la Figura 4.7 se muestran las respuestas de diez estudiantes a cada subtitulado. Se observa que las respuestas no son independientes ya que cada estudiante tiene un preferencia por uno o varios niveles de respuesta en cada test. Por otro lado, los ítems no son independientes unos de otros ya que pretenden medir la misma variable subyacente. Además, el interés no es conocer el valor concreto de sus coeficientes sino su valor en relación a los coeficientes de los otros ítems. En Bürkner (2021a, pp. 14-16) y en Bürkner y Vuorre (2019, pp. 19-20) se puede encontrar un ejemplo con esta parametrización aplicada a una escala de Likert.

Modelo Response ~ Treat * Period + (1 | Subject)

El primer modelo que se propone es un modelo que mantiene los predictores `Treat` y `Period` y su interacción (equivalente al efecto secuencia) como efectos fijos que fueron seleccionadas en la sección anterior (ver Sección 4.2) e incorpora los estudiantes como efectos aleatorios sobre los interceptos:

$$Nivel\ 1 : \text{logit}(P(Response_{ij} \leq k)) = \tau_{kj} - \beta_1 \text{Treat}_{ij} - \beta_2 \text{Period}_{ij} - \beta_3 \text{Treat}_{ij} : \text{Period}_{ij}$$
$$Nivel\ 2 : \tau_{kj} = \tau_k + Subject_{0j}$$

donde ij es la observación i del estudiante j. Obsérvese que ahora los interceptos τ_{kj} se descomponen en una parte fija y común para cada nivel de respuesta k, τ_k y una parte variable específica para cada estudiante $Subject_{0j}$. Para ajustar el

modelo se va a utilizar la función `clmm` del paquete `ordinal` (ver Christensen 2022) ya que permite la inclusión de efectos aleatorios.

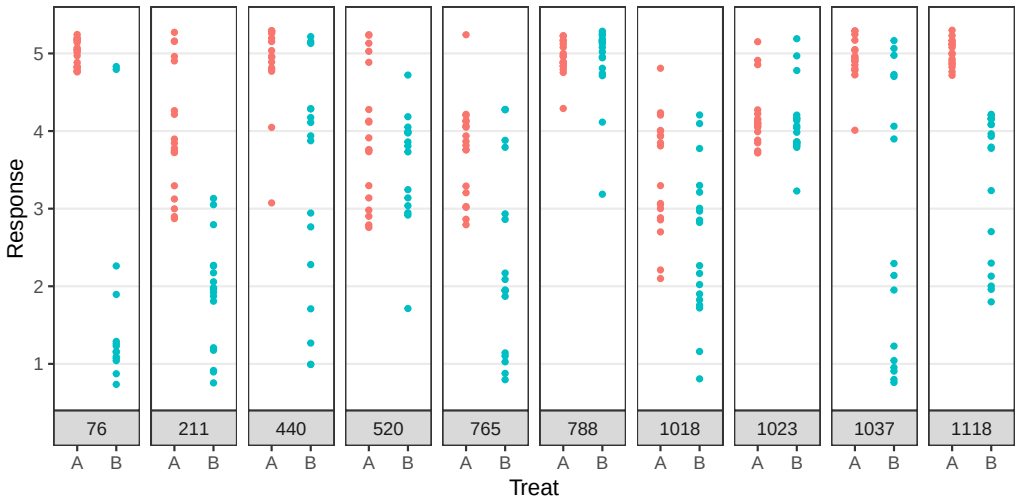

Figura 4.7: Respuestas de los estudiantes por nivel de subtitulado.

```r
options(contrasts = rep("contr.sum", 2))
clmm_treat.period_subject <- clmm(
    Response ~ Treat * Period + (1 | Subject),
    data = df_response
)
summary(clmm_treat.period_subject)
```

```
Cumulative Link Mixed Model fitted with the Laplace approximation

formula: Response ~ Treat * Period + (1 | Subject)
data:    df_response

 link  threshold nobs logLik   AIC     niter     max.grad cond.H
 logit flexible  2980 -3655.71 7327.41 765(3046) 1.63e-03 8.1e+01

Random effects:
 Groups  Name        Variance Std.Dev.
 Subject (Intercept) 1.278    1.131
Number of groups:  Subject 87

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
Treat1         1.05368    0.03999  26.346  < 2e-16 ***
Period1        0.15662    0.03604   4.346 1.39e-05 ***
Treat1:Period1 0.14262    0.12677   1.125    0.261
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
    Estimate Std. Error z value
1|2  -3.7046     0.1523 -24.332
2|3  -2.0298     0.1349 -15.050
3|4  -1.1012     0.1310  -8.406
4|5   0.8281     0.1299   6.375
```

En la parte de efectos fijos: los interceptos tienen valores similares al modelo de efectos fijos (ver Sección 4.2) aunque los coeficientes incrementan ligeramente

su valor. Esto indica una mayor distancia entre las respuestas de los subtitulados *A* y *B*. En este modelo el efecto secuencia no es significativo. En cuanto a los efectos aleatorios: la varianza del intercepto aleatorio de los estudiantes es 1.28. En la Figura 4.8 se muestran los valores de los interceptos estimados de los estudiantes. La media de estos interceptos como se espera es cercana a cero (-0.008).

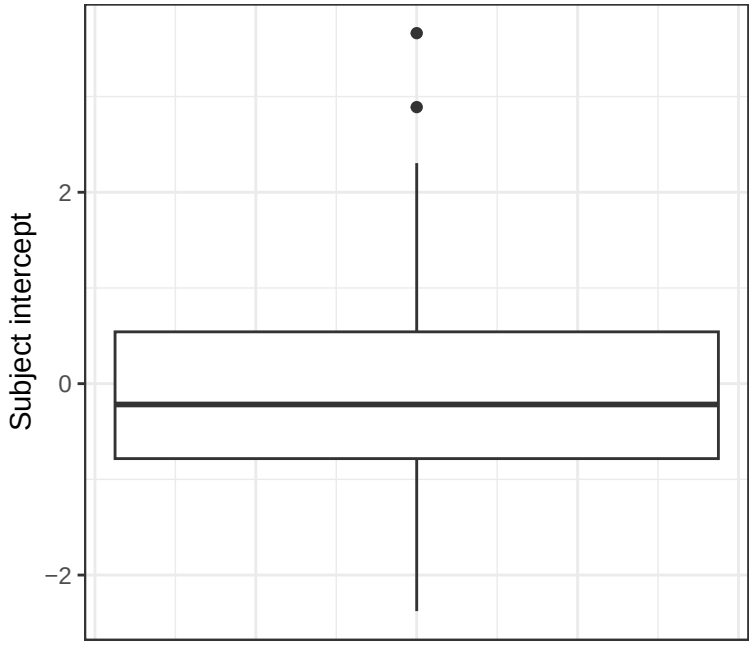

Figura 4.8: Distribución de interceptos aleatorios por estudiante en el modelo Response ~ Treat * Period + (1 | Subject)

Modelo Response ~ Treat * Period + (1 + Treat | Subject)

Es posible que cada estudiante valore con diferente criterio cada subtitulado. Para estimarlo, se propone el siguiente modelo:

$$Nivel\ 1: \text{logit}(P(Response_{ij} \leq k)) = \tau_{kj} - \beta_{1j}\text{Treat}_{ij} - \beta_2\text{Period}_{ij} - \beta_3\text{Treat}_{ij} * \text{Period}_{ij}$$

$$Nivel\ 2: \tau_{kj} = \tau_k + Subject_{0j}$$

$$\beta_{1j} = \beta_1 + Subject_{1j}$$

Ahora el parámetro β_{1j} del subtitulado tiene dos componentes: Uno común a todos los niveles de respuesta β_1 y otro particular de cada estudiante $Subject_{1j}$. El modelo ajustado ocasiona que solo `Treat1` sea significativo, ya que ni el periodo ni la secuencia lo son. En los efectos aleatorios la correlación entre intercepto y pendiente es prácticamente nula.

```
options(contrasts = rep("contr.sum", 2))
clmm_treat.period_treat.subject <- clmm(
    Response ~ Treat * Period + (1 + Treat | Subject),
```

```
    data = df_response
)
summary(clmm_treat.period_treat.subject)
```

```
Cumulative Link Mixed Model fitted with the Laplace approximation

formula: Response ~ Treat * Period + (1 + Treat | Subject)
data:    df_response

 link  threshold nobs logLik   AIC     niter     max.grad cond.H
 logit flexible  2980 -3429.88 6879.76 905(6264) 1.33e-03 8.2e+01

Random effects:
 Groups  Name        Variance Std.Dev. Corr
 Subject (Intercept) 1.712    1.308
         Treat1      1.042    1.021    -0.062
Number of groups:  Subject 87

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
Treat1          1.2938     0.1197  10.809   <2e-16 ***
Period1         0.1620     0.1171   1.383    0.167
Treat1:Period1  0.1327     0.1464   0.906    0.365
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
    Estimate Std. Error z value
1|2  -4.2633     0.1761 -24.210
2|3  -2.3321     0.1562 -14.932
3|4  -1.2656     0.1520  -8.324
4|5   0.9659     0.1509   6.400
```

Comparación de modelos

Se pueden comparar los modelos con el test de razón de verosimilitud que se realiza con la función `anova` del paquete `ordinal` (ver Christensen 2022). Se comprueba que en este test resulta significativamente mejor el último modelo:

```
anova(clmm_treat.period_subject, clmm_treat.period_treat.subject)
```

```
Likelihood ratio tests of cumulative link models:

                               formula:
clmm_treat.period_subject      Response ~ Treat * Period + (1 | Subject)
clmm_treat.period_treat.subject Response ~ Treat * Period + (1 + Treat | Subject)
                               link: threshold:
clmm_treat.period_subject      logit flexible
clmm_treat.period_treat.subject logit flexible

                               no.par    AIC  logLik LR.stat df Pr(>Chisq)
clmm_treat.period_subject           8 7327.4 -3655.7
clmm_treat.period_treat.subject    10 6879.8 -3429.9  451.66  2  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Elección del mejor modelo

En el apartado anterior se introdujo a los estudiantes como efecto aleatorio. Como se ha dicho, los ítems también pueden modelizarse como aleatorios. Esto produce una multiplicidad de modelos. Los siguientes son los que se han comparado:

- Response ~ (1 | Subject)
- Response ~ (1 + Treat | Subject)
- Response ~ (1 + Treat | Item)
- Response ~ Treat + (1 + Treat | Subject)
- Response ~ Treat + (1 + Treat | Item)
- Response ~ Treat*Period + (1 + Treat | Subject)
- Response ~ Treat*Period + (1 + Treat | Item)
- Response ~ Treat*Period + (1 + Period | Subject) + (1 + Treat | Item)
- Response ~ Treat + (1 + Treat | Subject) + (1 + Treat | Item)
- Response ~ Treat*Period + (1 + Treat | Subject) + (1 + Treat | Item)

El último de ellos produce un resultado significativo en el test de razón de verosimilitud con todos los demás. Sin embargo los parámetros de todos los modelos tienen valores similares por lo que no cambia la interpretación que se haga de ellos en cada modelo. Este modelo tiene un AIC menor que los modelos ordinales ajustados en el apartado anterior (ver Sección 4.2) incluso si a esos modelos se les añade como factor predictor `Item`. Será este, por lo tanto, el modelo seleccionado.

La Ecuación 4.2 del modelo seleccionado es la siguiente:

$$\begin{aligned} &Nivel\ 1: \text{logit}(P(Response_{ijl} \leq k)) = \tau_{kjl} - \beta_{1jl}\text{Treat}_{ijl} - \beta_2\text{Period}_{ijl} - \beta_3\text{Treat}_{ijl} : \text{Period}_{ijl} \\ &Nivel\ 2: \tau_{kjl} = \tau_k + Subject_{0j} + Item_{0l} \\ &\qquad \beta_{1jl} = \beta_1 + Subject_{1j} + Item_{1l} \end{aligned} \tag{4.2}$$

donde ijl se corresponde con la observación i-ésima del estudiante j e ítem l. Ahora los interceptos y el coeficiente del subtitulado se componen de tres sumandos: una parte fija, una parte que depende del estudiante y una parte que depende del ítem de Likert.

El resumen de parámetros del modelo ajustado es el siguiente:

```
options(contrasts = rep("contr.sum", 2))
clmm_treat.period.subject.item <- clmm(
    Response ~ Treat * Period + (1 + Treat | Subject) + (1 + Treat | Item),
    data = df_response
)

summary(clmm_treat.period.subject.item)
```

```
Cumulative Link Mixed Model fitted with the Laplace approximation
```

```
formula: Response ~ Treat * Period + (1 + Treat | Subject) + (1 + Treat |
    Item)
data:    df_response

 link  threshold nobs logLik   AIC     niter       max.grad cond.H
 logit flexible  2980 -3186.06 6398.11 1468(12273) 2.37e-03 1.5e+02

Random effects:
 Groups  Name        Variance Std.Dev. Corr
 Subject (Intercept) 2.2176   1.4892
         Treat1      1.3650   1.1683   -0.128
 Item    (Intercept) 0.4831   0.6950
         Treat1      0.4655   0.6823   -0.528
Number of groups:  Subject 87,  Item 18

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
Treat1          1.4320     0.2102   6.811  9.7e-12 ***
Period1         0.1730     0.1325   1.306    0.192
Treat1:Period1  0.1397     0.1654   0.845    0.398
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
    Estimate Std. Error z value
1|2  -5.0033     0.2619 -19.104
2|3  -2.6499     0.2417 -10.962
3|4  -1.3659     0.2376  -5.748
4|5   1.1837     0.2365   5.006
```

Con este modelo los efectos secuencia y periodo no son significativos. En cualquier caso, se mantienen ya que el test de razón de verosimilitud resulta significativo en este modelo respecto al modelo sin estos predictores.

Modelado bayesiano

Existen muchos paquetes en R para hacer inferencia bayesiana. Algunos de los más populares son:

- OpenBUGS y WinBUGS: basado en el muestreo de Gibbs.
- JAGS: también utiliza el muestreo de Gibbs.
- Stan: Más moderno y con una comunidad de desarrollo más activa que los anteriores. Utiliza muestreo HMC (Hamiltonian Monte Carlo) y NUTS (no U-turn sampler). Stan tiene un lenguaje similar a C para definir modelos aunque hay muchos paquetes basados en Stan que facilitan la especificación de modelos con una sintaxis más sencilla. En este trabajo se utilizará uno de ellos, `brms` (ver Bürkner 2021b) . La sintaxis de especificación de modelos con este paquete es idéntica a la que se ha utilizado en la sección anterior.
- INLA: Evita la simulación *MCMC* haciendo más rápida la convergencia. Es menos flexible ya que solo se pueden especificar modelos de la familia exponencial.

Se han comparado múltiples modelos usando la función `LOO` que realiza una validación cruzada bayesiana `leave-one-out` similar a la que se explicó en la Sección 2.4. El mejor modelo ha resultado ser el mismo que se seleccionó en modelos mixtos (ver Ecuación 4.2). Es decir:

Response ~ Treat*Period + (1 + Treat | Subject) + (1 + Treat | Item)

```
options(contrasts = rep("contr.sum", 2))
brm_treat.period.subject.item <- brm(
    Response ~ Treat * Period + (1 + Treat | Subject) + (1 + Treat | Item),
    data = df_response,
    family = cumulative("logit"),
    iter = 4000,
    sample_prior = TRUE,
    file = "models/brm_treat.period.subject.item",
    file_refit = "on_change"
)
```

El modelo utiliza como factores con efectos fijos (`complete pooling` en terminología bayesiana) el nivel de subtitulado y el periodo y la interacción entre ambos; y como efectos aleatorios (`partial pooling`) los sujetos y los ítems del test, cada uno de ellos con un intercepto y un nivel de subtitulado variable. El resumen del modelo es el siguiente:

```
summary(brm_treat.period.subject.item)
```

```
 Family: cumulative
  Links: mu = logit; disc = identity
Formula: Response ~ Treat * Period + (1 + Treat | Subject) + (1 + Treat | Item)
   Data: df_response (Number of observations: 2980)
  Draws: 4 chains, each with iter = 4000; warmup = 2000; thin = 1;
         total post-warmup draws = 8000

Group-Level Effects:
~Item (Number of levels: 18)
                    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
sd(Intercept)           0.77      0.16     0.53     1.15 1.00     1468
sd(Treat1)              0.77      0.15     0.53     1.12 1.00     1935
cor(Intercept,Treat1)  -0.46      0.20    -0.78     0.01 1.00     1421
                    Tail_ESS
sd(Intercept)           2605
sd(Treat1)              3580
cor(Intercept,Treat1)   2555

~Subject (Number of levels: 87)
                    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
sd(Intercept)           1.54      0.14     1.29     1.85 1.00     1484
sd(Treat1)              1.21      0.11     1.01     1.45 1.00     1519
cor(Intercept,Treat1)  -0.11      0.12    -0.34     0.14 1.00     1228
                    Tail_ESS
sd(Intercept)           2741
sd(Treat1)              3081
cor(Intercept,Treat1)   2475

Population-Level Effects:
             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept[1]    -4.95      0.28    -5.52    -4.42 1.00      785     1753
Intercept[2]    -2.59      0.26    -3.11    -2.09 1.01      743     1636
```

```
Intercept[3]       -1.30      0.26    -1.82    -0.82 1.01      726      1532
Intercept[4]        1.25      0.26     0.74     1.75 1.00      745      1571
Treat1              1.46      0.23     1.01     1.92 1.00     1046      2071
Period1             0.17      0.14    -0.09     0.44 1.00      848      1714
Treat1:Period1      0.14      0.17    -0.20     0.47 1.01      686      1127

Family Specific Parameters:
     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
disc     1.00      0.00     1.00     1.00   NA       NA       NA

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Se han mantenido las distribuciones de probabilidad a priori que por defecto utiliza `brm` confiando en que sus parámetros son adecuados. Sin embargo, conviene comprobar que realmente sea así. En la Tabla 4.12 se muestran las distribuciones a priori de los parámetros aleatorios del modelo. En la Figura 4.9 se constata que toman valores razonables y no informativos.

Tabla 4.12: Distribuciones a priori del modelo ordinal seleccionado.

prior	class	coef	group	resp	dpar	nlpar	lb	ub	source
	b								default
	b	Period1							default
	b	Treat1							default
	b	Treat1:Period1							default
student_t(3, 0, 2.5)	Intercept								default
student_t(3, 0, 2.5)	Intercept	1							default
student_t(3, 0, 2.5)	Intercept	2							default
student_t(3, 0, 2.5)	Intercept	3							default
student_t(3, 0, 2.5)	Intercept	4							default
lkj_corr_cholesky(1)	L								default
lkj_corr_cholesky(1)	L		Item						default
lkj_corr_cholesky(1)	L		Subject						default
student_t(3, 0, 2.5)	sd						0		default
student_t(3, 0, 2.5)	sd		Item						default
student_t(3, 0, 2.5)	sd	Intercept	Item						default
student_t(3, 0, 2.5)	sd	Treat1	Item						default
student_t(3, 0, 2.5)	sd		Subject						default
student_t(3, 0, 2.5)	sd	Intercept	Subject						default
student_t(3, 0, 2.5)	sd	Treat1	Subject						default

Es importante asegurar que el entrenamiento ha convergido a su distribución a posteriori. En la tabla de resumen se constata que el valor de `Rhat` [8] es inferior a 1.1 y el de `ESS` [9] superior a 400 en todos los parámetros, que son umbrales que no se deberían violar (ver Bürkner y Vuorre 2019). En la Figura 4.10 se comprueba que las cadenas $MCMC$ de muestreo de la distribución a posteriori se mezclan correctamente y no se aprecia autocorrelación en ninguno de los parámetros. Por último, en la Figura 4.11 se muestra una comparación entre los histogramas construidos con los datos de las respuestas a los test con los intervalos de credibilidad marginales de la función predictiva a posteriori del modelo. En la mayoría de los

[8] `Rhat` es una medida utilizada para evaluar la convergencia de las Cadenas de Markov Monte Carlo ($MCMC$) en el muestreo bayesiano. Compara la varianza de cada cadena individual de $MCMC$ con la varianza entre diferentes cadenas. Si las cadenas convergen, se espera que sus valores sean similares y, por lo tanto, el valor de `Rhat` será próximo a 1.

[9] `ESS` (Efficient Sample Size) es una estimación del número de muestras independientes obtenidas en el muestreo $MCMC$.

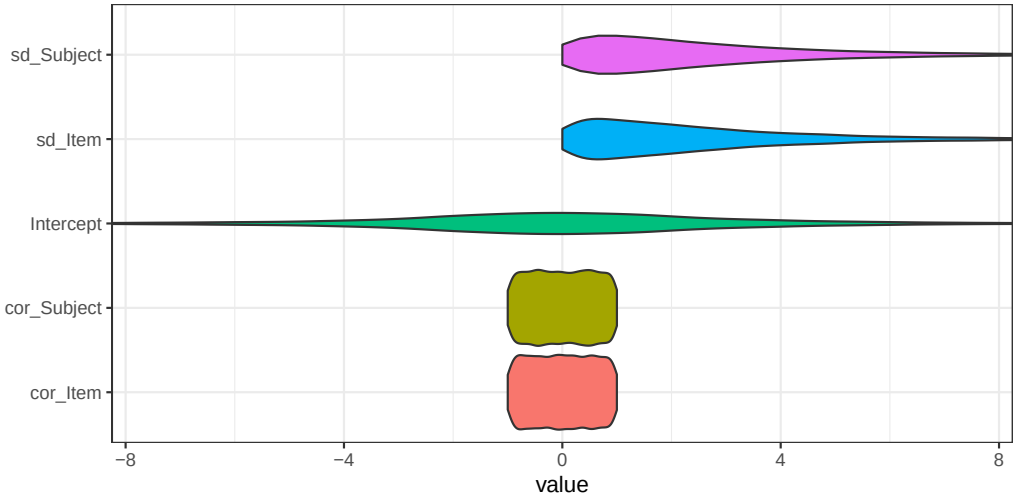

Figura 4.9: Distribuciones a priori del modelo Response ~ Treat * Period + (1 + Treat | Subject) + (1 + Treat | Item).

ítems el muestreo se ajusta bastante bien al histograma de respuestas; aunque en algunos ítems, como el `Q16` o el `Q17`, se aprecian diferencias relevantes.

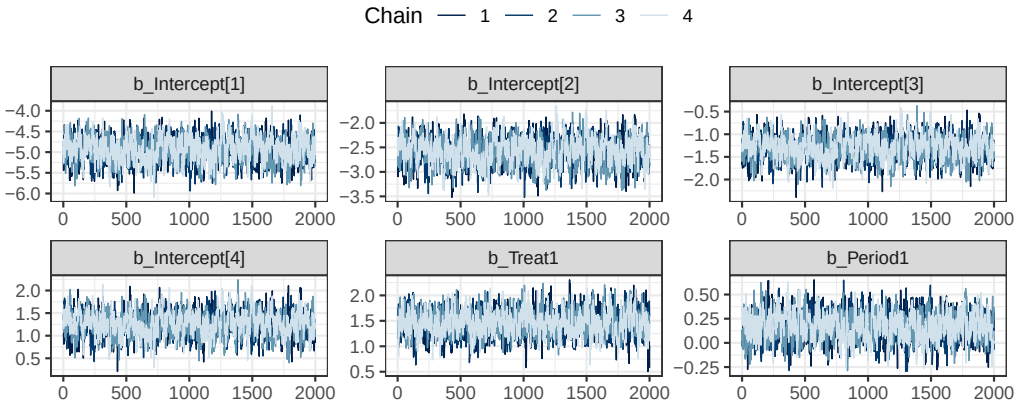

Figura 4.10: Cadenas MCMC del modelo Response ~ Treat * Period + (1 + Treat | Subject) + (1 + Treat | Item).

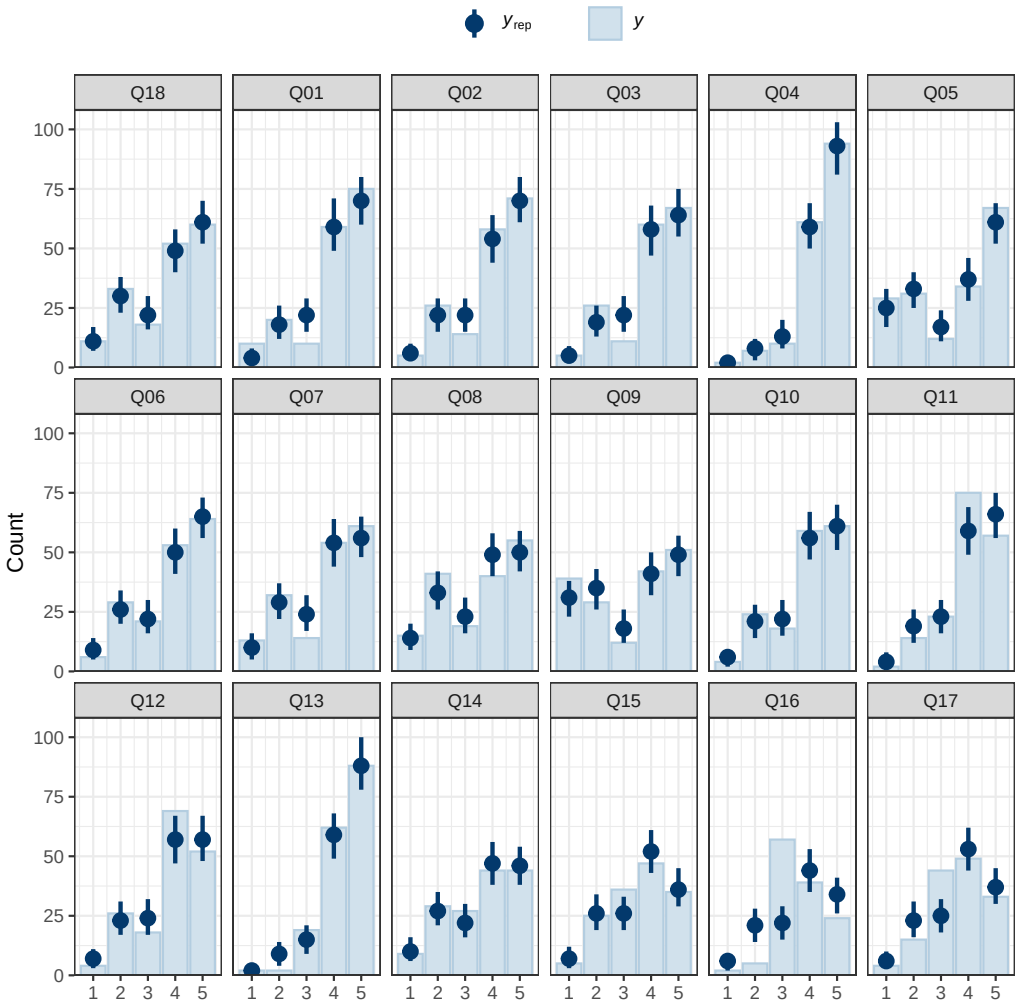

Figura 4.11: Comparación de los valores reales con los obtenidos a partir de la función predictiva a posteriori del modelo Response ~ Treat * Period + (1 + Treat | Subject) + (1 + Treat | Item).

Capítulo 5

Resultados

En el Capítulo 4 se realizó una exploración de los datos y se adecuaron los modelos presentados en el Capítulo 2 al diseño del experimento del subtitulado. En este capítulo se comentan los resultados de los modelos seleccionados siguiendo el mismo orden expositivo, comenzando por el análisis del OR, continuando por la Regresión Logística y finalizando con la Regresión Ordinal.

Comparación con *Odds Ratio*

El contraste de hipótesis del *log OR* del nivel subtitulado y secuencia (ver Sección 4.2) no produce significación estadística en ningún nivel de respuesta por lo que, según esta prueba estadística, el orden en el que se ven los vídeos no influye en la respuesta de los estudiantes (ver Tabla 5.1).

Tabla 5.1: Log OR ~ Treat + Seq + Response

Response	Estimate	Std. Error	z value	Pr(>|z|)
No sé / No contesto	0.190	0.327	0.580	0.562
Muy en desacuerdo	-0.135	1.012	-0.134	0.894
En desacuerdo	-0.244	0.291	-0.838	0.402
Neutral	0.152	0.214	0.711	0.477
De acuerdo	-0.210	0.134	-1.570	0.116
Muy de acuerdo	0.117	0.137	0.855	0.393

Sin embargo, si se realiza este contraste entre subtítulos y periodos, se constata la existencia de un efecto periodo de signo contrario para los ítems 4 y 5 (ver Tabla 5.2). El test es significativo porque el ratio entre subtítulos de respuestas con valor 4 es diferente en cada periodo habiendo mayor cantidad de respuestas 4 en el segundo periodo que en el primero. Con las respuestas 5 ocurre lo contrario: la proporción es mayor en el primer periodo. La Figura 5.1 permite una

comprobación visual. Esto indica que los estudiantes de ambos grupos prestaron más atención o fueron más exigentes en el segundo visionado y valoraron relativamente peor el segundo vídeo. Que el efecto periodo sea de signo contrario en dos respuestas no debe sorprender en este diseño de experimento, ya que un test es un juego de suma cero: la valoraciones que se ganan o se pierden en un nivel de respuesta necesariamente se reparten entre el resto de niveles. En cualquier caso, el efecto periodo es cuantitativa y cualitativamente pequeño. Al afectar solo al intercambio de valoraciones entre los niveles 4 y 5, es simplemente una pequeña corrección en la valoración del subtitulado y cualitativamente es poco importante ya que las respuestas 4 y 5 son ambas valoraciones positivas.

Tabla 5.2: Log OR ~ Treat + Period + Response

Response	Estimate	Std. Error	z value	Pr(>|z|)
No sé / No contesto	0.335	0.327	1.022	0.307
Muy en desacuerdo	0.135	1.012	0.134	0.894
En desacuerdo	-0.121	0.291	-0.416	0.677
Neutral	0.055	0.214	0.259	0.796
De acuerdo	-0.851	0.134	-6.367	0.000
Muy de acuerdo	0.486	0.137	3.557	0.000

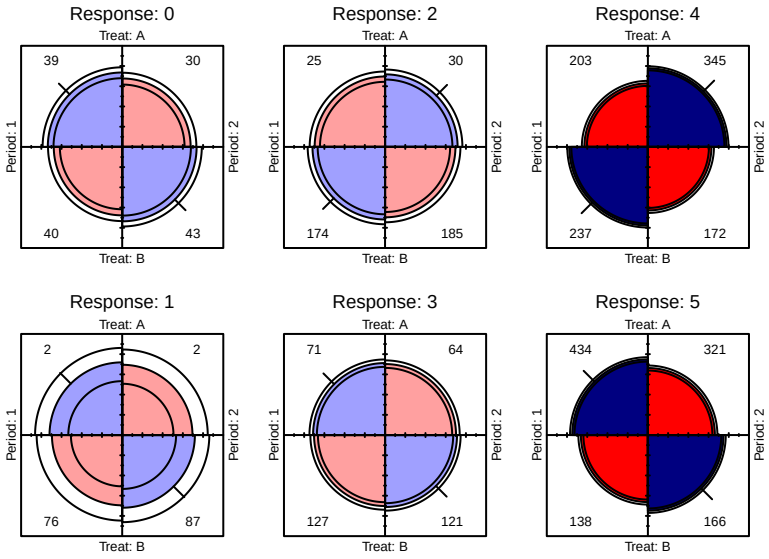

Figura 5.1: OR ~ Treat + Period + Response

Modelado

Regresión Logística

En la Sección 4.2 se explicó una forma de crear una variable dicotómica que permite ajustar los datos a una Regresión Logística. Concretamente, se creó la variable respuesta *Improve* que compara si las respuestas de cada estudiante a cada Ítem entre los niveles de subtitulado (*A* frente *B*) han mejorado (valor 1)

Tabla 5.3: Resumen de los modelos de Regresión Logística.

	Improve (A>B)	Level == 'Positivo'
(Intercept)	0.465	1.474***
	(0.295)	(0.284)
Treat1		1.548***
		(0.223)
SD (Intercept Subject)	1.703	1.590
SD (Treat1 Subject)		1.082
Cor (Intercept~Treat1 Subject)		−0.093
SD (Intercept Item)	0.931	0.893
SD (Treat1 Item)		0.726
Cor (Intercept~Treat1 Item)		−0.119
Num.Obs.	1451	2980
AIC	1553.6	2390.0
BIC	1569.5	2438.0

o se han mantenido o empeorado (valor 0). El modelo que se propone en esta sección no tiene en cuenta el efecto secuencia porque resultó no significativo en el análisis y, en cambio, se incluye como efectos aleatorios los estudiantes y los ítems sobre el intercepto ya que, como se ha explicado, son variables que no se pueden considerar independientes:

$$\text{Improve} \sim 1 + (1 \mid \text{Subject}) + (1 \mid \text{Item})$$

```
glmer_improve_subject_question <- glmer(
    Improve ~ 1 + (1 | Subject) + (1 | Item),
    family = "binomial", data = df_improve
)
```

El resumen del modelo ajustado con la función `glmer` del paquete `lme4` (ver Bates et al. 2015) produce los resultados de la columna izquierda de la Tabla 5.3. El intercepto del modelo ajustado es 0.465 (std.error 0.295). Por ello, la probabilidad de que se otorgue una mayor puntuación en *A* que en *B* es del 61.42%. La proporción de varianza explicada por los efectos aleatorios (*ICC*) es 0.534.

Una forma alternativa de Regresión Logística es simplemente saber si cada respuesta es positiva (4 ó 5) frente a si es negativa o neutra (1, 2, ó 3). Como efecto fijo se incorpora el nivel de subtitulado y como efectos aleatorios el estudiante y el ítem ambos con intercepto y pendiente sobre el subtitulado variables por tener mejor valor de AIC que otros modelos probados [1]. La fórmula del modelo es la siguiente:

$$\text{I(Level == «Positivo»)} \sim \text{Treat} + (1 + \text{Treat} \mid \text{Subject}) + (1 + \text{Treat} \mid \text{Item})$$

[1] Por problemas de convergencia, no se ha podido ajustar el modelo I(Level == «Positivo») ~ Treat*Period + (1 + Treat | Subject) + (1 + Treat | Item)

En la columna derecha de la Tabla 5.3 se muestra el resumen del modelo. En este caso, el intercepto tiene una valor de 1.474. Por lo que la probabilidad de respuesta positiva en cualquier ítem y nivel de subtitulado es 81.4%. El coeficiente `Treat1` es significativo y tiene valor 1.548 y es el valor en `logits` que se añade o se quita en función de si el subtitulado es el *A* o el *B*. Esto se traduce en que la posibilidad de una respuesta positiva en el subtitulado *A* es 95.4%. En el subtitulado *B* esta probabilidad se reduce a 48.2%. Por último, el valor de ICC del modelo es 0.604

Regresión Ordinal

En la Sección 4.2 se evaluaron distintas parametrizaciones de la Regresión Ordinal Acumulativa tanto desde el punto de vista frecuentista como bayesiano, considerando únicamente efectos fijos y también efectos aleatorios. Finalmente, tanto en el análisis frecuentista como en el bayesiano, el modelo que resultó ser más parsimonioso (evaluado con *LRT* y con AIC) es el de la Ecuación 4.2, que se reproduce aquí en sintaxis R:

```
Response ~ Treat*Period + (1 + Treat | Subject) + (1 + Treat | Item)
```

Este modelo incluye como efectos fijos el nivel de subtitulado (`Treat`), el periodo (`Period`) y su interacción; y como efectos aleatorios el estudiante (`Subject`) y el ítem (`Item`). Ambos con interceptos y pendientes variables por nivel de subtítulo. Los coeficientes estimados son muy similares tanto en el paradigma frecuentista como en el bayesiano. En la Tabla 5.4 se comparan las estimaciones producidas por ambos modelos. Como ya se dijo, el efecto más importante es el debido al subtitulado (coeficiente frecuentista 1.432). En comparación con él, los efectos debido al periodo y la secuencia son muy pequeños y no significativos (coeficientes 0.173 y 0.14 respectivamente). La proporción de la varianza explicada debida a efectos aleatorios (ICC) es 0.579.

En la Figura 5.2 se muestran las predicciones del modelo por nivel de subtítulo y periodo. El modelo predice para el subtitulado *B* el nivel 4 de respuesta como el más probable seguido del 3, mientras para el subtitulado *A* el nivel de respuesta más probable es el 5 seguido del 4. En el subtitulado *B* apenas hay diferencias entre periodos, sin embargo, en el subtitulado *A* hay mayor probabilidad del nivel de respuesta 5 en el periodo 1 y nivel de respuesta 4 en el periodo 2.

Tabla 5.4: Comparación frecuentista/bayesiano de coeficientes estimados en el modelo ordinal.

	ordinal::clmm			brms::brm		
Name	Est.	conf.2.5%	conf.97.5%	Est.	cred.2.5%	cred.97.5%
1|2	-5.00	-5.52	-4.49	-4.94	-5.52	-4.42
2|3	-2.65	-3.12	-2.18	-2.58	-3.11	-2.09
3|4	-1.37	-1.83	-0.90	-1.30	-1.82	-0.82
4|5	1.18	0.72	1.65	1.25	0.74	1.75
Treat1	1.43	1.02	1.84	1.46	1.01	1.92
Period1	0.17	-0.09	0.43	0.17	-0.09	0.44

Treat1:Period1	0.14	-0.18	0.46	0.14	-0.20	0.47
Item.sd(Intercept)	0.70			0.75	0.53	1.15
Item.sd(Treat1)	0.68			0.75	0.53	1.12
Subject.sd(Intercept)	1.49			1.53	1.29	1.85
Subject.sd(Treat1)	1.17			1.21	1.01	1.45
Item.cor(Intercept,Treat1)	-0.53			-0.49	-0.78	0.01
Subject.cor(Intercept,Treat1)	-0.13			-0.11	-0.34	0.14

En la Figura 5.3 se representan 50 muestras de la esperanza de la distribución predictiva a posteriori para cada ítem y nivel de subtitulado marginalizados por periodo y estudiante. La primera conclusión que se puede extraer es que el modelo tiene bastante incertidumbre sobre los valores de respuesta a cada ítem no superando casi nunca el 50% de probabilidad para todos los ítems y niveles de subtitulado. En general se observa en la mayoría de los ítems del nivel de subtitulado *A* que los alumnos están bastante seguros de que la respuesta a los ítems debe ser 4 ó 5, asignando una muy baja probabilidad a los valores 1, 2, ó 3, pero habiendo bastante incertidumbre respecto cuál de los dos valores (4 ó 5) asignar. En el nivel de subtitulado *B* la situación es bastante más confusa. Aunque la opción de respuesta preferida es 4 y las menos preferidas son la 5 y la 1, hay bastante mezcla entre las opciones de respuesta 2, 3 y 4. En cuanto al análisis individualizado por ítem se llega a las siguientes conclusiones:

- En los ítems *Q*04 y *Q*13 los estudiantes no aprecian defectos en el subtitulado ni diferencias entre un nivel y otro. Son valoradas en ambos subtitulados con puntuaciones de 4 y de 5.

- En los ítems *Q*15, *Q*16 y *Q*17, la opción de respuesta más probable es 4. El modelo asigna una baja probabilidad de respuesta a la opción 1 y similares al resto. La probabilidad de la opción 5 decrece ligeramente entre subtitulado *A* y *B* y lo contrario ocurre con las opciones 2 y 3.

- Las muestras de los ítems *Q*01, *Q*02, *Q*03, *Q*10, *Q*11 y *Q*12 son similares a las anteriores. Particularmente en lo referente a que la respuesta más probable en el subtitulado *B* es 4. En el subtitulado *A* hay preferencia por 4 y 5. El nivel 5 cae acusadamente en el subtitulado *B* y en este nivel aumenta ligeramente la probabilidad de respuesta 2 y 3.

- Los ítems *Q*06, *Q*07, *Q*14 y *Q*18 tampoco son muy diferentes de los anteriores. En general el modelo predice mayor probabilidad de respuesta para 5 en el subtitulado *A* pero este valor es con alta probabilidad cercano a cero en el subtitulado *B*. En el subtitulado *B* la probabilidad de respuesta 2, 3 ó 4 es similar.

- Los ítems *Q*05, *Q*08 y *Q*09 son los que más diferencias entre subtitulados presentan. La respuesta más probable en el subtitulado *A* es 5 (en *Q*08 y en *Q*09 muy parecida a 4). Por contra, en el subtitulado *B* las respuestas 4 y 5 tienden a cero, siendo la más probable la respuesta 2. En los ítems *Q*05 y *Q*09 la segunda respuesta más probable al subtitulado *B* es 1 y 4 en el ítem *Q*08.

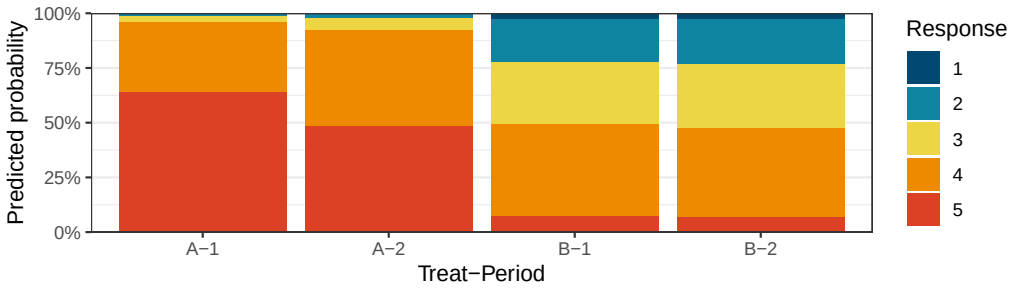

Figura 5.2: Probabilidades de respuesta para el modelo ordinal Response ~ Treat * Period + (1 + Treat | Subject) + (1 + Treat | Item)

En definitiva, el modelo predice que los estudiantes están bastante de acuerdo en que en los ítems $Q05$ y $Q09$ hay una diferencia de calidad importante entre subtitulados. También están de acuerdo en que en los ítems $Q04$ y $Q13$ no hay apenas cambio entre los subtitulados. En los ítems $Q15$, $Q16$ y $Q17$ hay una gran confusión en ambos niveles de subtitulado predominando la respuesta 4 y siendo muy parecidas las respuestas en ambos niveles. En el resto la confusión se circunscribe al nivel de subtitulado B, ya que en el nivel A las opciones 4 y 5 predominan.

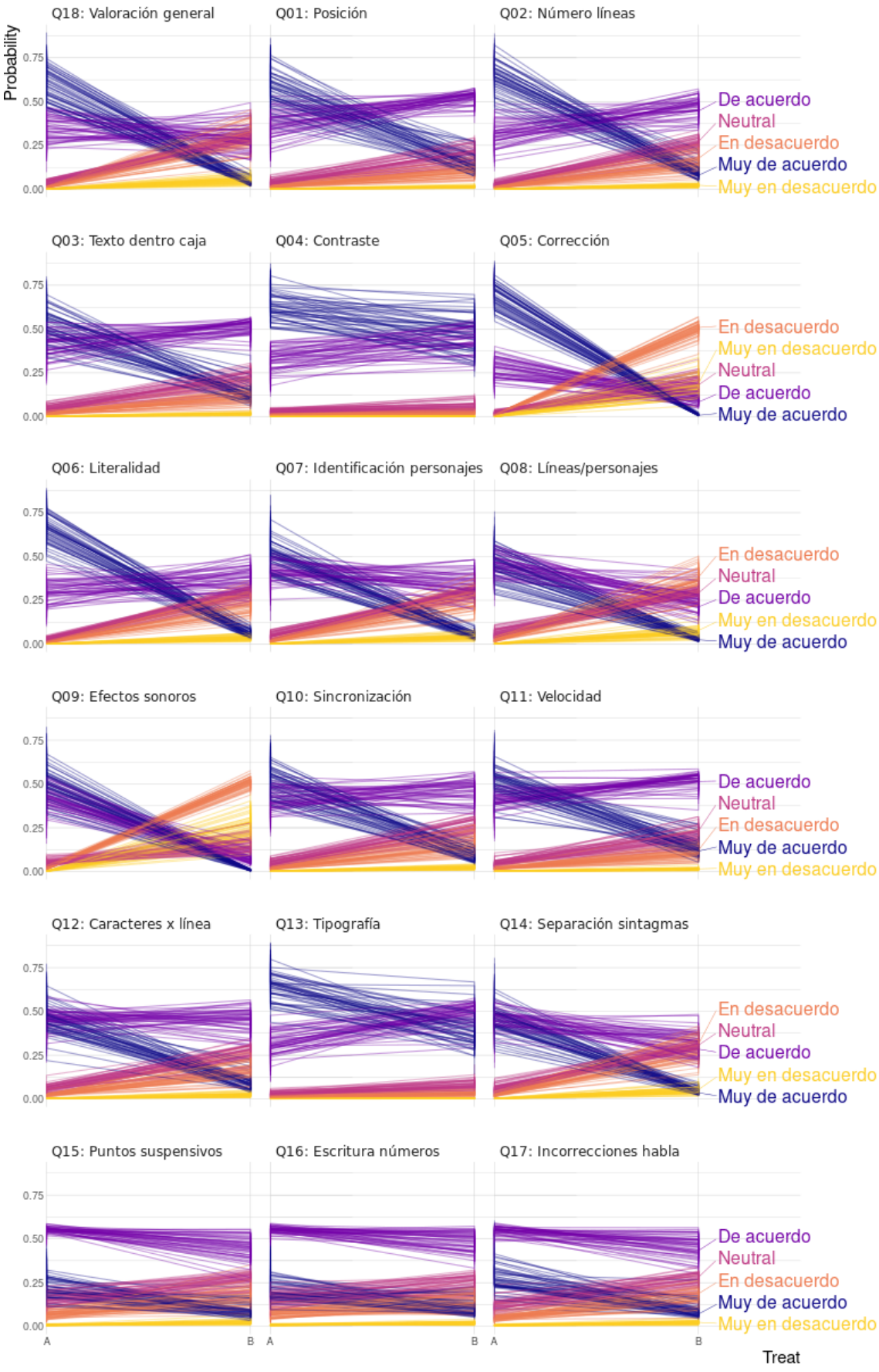

Figura 5.3: Muestreo de la función predictiva a posteriori por tratamiento e ítem.

Capítulo 6

Discusión

En este capítulo se discuten los resultados una vez «descubierto el ciego» y se responde a la pregunta de investigación y a los objetivos específicos planteados en la Sección 1.2. Las respuestas a estas preguntas llevan a concluir que se ha cumplido el objetivo principal del estudio: los estudiantes del curso MOOC son capaces de evaluar las diferencias en la calidad del subtitulado de dos vídeos. Se discuten también las limitaciones del estudio y posibles mejoras al mismo.

Una vez realizado el análisis estadístico y obtenidos los resultados (ver Capítulo 5) se reveló que el subtitulado que en este trabajo se ha denominado *A* se corresponde con el vídeo correctamente subtitulado. Adicionalmente se ha suministrado un documento que contiene los errores introducidos en el subtitulado *B*. A partir del mismo, se ha elaborado la Tabla 6.1 en la que se muestra la correspondencia de cada error con los ítems de la escala de Likert a la que respondieron los estudiantes (ver en la Tabla 3.1 una descripción textual de cada ítem). Para 7 de los 18 ítems (Q01, Q03, Q04, Q13, Q15, Q16, Q17) no se ha encontrado una adscripción clara en el documento de errores y, por lo tanto, esos ítems sirven de control del test y sería esperable que en ellos las respuestas de los estudiantes fueran similares en ambos subtitulados.

6.1 Respuestas a las preguntas de investigación y a los objetivos específicos

Para responder a las preguntas de investigación se van a utilizar los hallazgos del Análisis Exploratorio (ver Sección 4.1) y los resultados de los tres modelos comentados en el Capítulo 5:

- Regresión Logística con variable respuesta `Improve`, que calcula la probabilidad de que la respuesta a un ítem mejore entre subtitulados $A > B$

Tabla 6.1: Correspondencia entre los errores introducidos en el subtitulado del vídeo B y los ítems de la escala de Likert.

Error nº	Subtítulo incorrecto	Requisito que se incumple	Ítems
1	Hola este video nos ba a servir	Los subtítulos deben ser correctos ortográfica y gramaticalmente.	Q05
2	para hacer una prácti ca de subtitulado	No se deben separar en dos líneas las sílabas de la misma palabra.	Q14
3	Podemos pensar en personas que no entienden bien un determinado idioma	Los subtítulos deben ocupar dos líneas y, excepcionalmente tres.	Q02
4	Muchos ejemplos que harán que estos subtítulos	Las conjunciones y los nexos deben ir en la línea inferior.	Q14
5	El texto del subtítulo es válido, pero su entrada debe producirse antes para que coincida con la información sonora.	Las entradas y salidas de los subtítulos deben coincidir con el movimiento labial, con la locución y/o con la información sonora.	Q10, Q09
6	Perdonad	Deben describirse los efectos sonoros que sean relevantes para la comprensión del vídeo.	Q09
7	¿Dígame? (EMI) Hola, soy Emilio.	Para cada participante en el diálogo debe comenzarse una nueva línea.	Q08
8	(ALE) Emilio, muy buenas. Mira, precisamente estoy grabando el vídeo	El máximo número de caracteres por línea es 37.	Q12
9	El texto del subtítulo es válido, pero su duración debe ser de al menos 3 segundos.	La velocidad de exposición del subtítulo debe permitir leerlo sin dificultad. La velocidad recomendada para los subtítulos es de unos 12 caracteres por segundo.	Q11
10	Luego voy a verte al despacho ¿Ok?	Los subtítulos deben ser literales.	Q06
11	Muy bien, estupendo. Aquí estaré. Hasta luego	Los diferentes personajes que intervienen en la obra audiovisual deben estar claramente identificados.	Q07

(ver Sección 5).

- Regresión Logística con variable respuesta `Positive`, que calcula la probabilidad de que la respuesta a un ítem sea positiva (4 ó 5).
- Regresión Ordinal con variable Respuesta `Response`, que calcula la probabilidad de cada nivel de respuesta (ver Sección 5).

Para mayor comodidad del lector se vuelven a plantear aquí la pregunta de investigación y los objetivos específicos:

> **Pregunta de investigación**
>
> ¿Son los estudiantes de un curso de creación de materiales accesibles capaces de evaluar las diferencias en la calidad del subtitulado de un vídeo?

El subtitulado *A*, que es el correcto, ha sido mejor evaluado por los estudiantes. Esto se ha constatado tanto en la exploración inicial como en cada uno de los tres modelos propuestos. Por ejemplo, en la exploración inicial se vio que la respuesta más frecuente en el subtitulado *A* es 5 y en el *B* es 4 y en el modelo con variable respuesta `Improve` predice que la probabilidad de que se otorgue una mayor puntuación en *A* que en *B* es 61.42%. Por lo tanto, se concluye respondiendo afirmativamente a la pregunta: los estudiantes del curso han sabido evaluar las diferencias en el subtitulado de los vídeos.

> Objetivo específico
>
> ¿En qué pautas de subtitulado los estudiantes tienen mayor **facilidad** para reconocer diferencias entre un subtitulado correcto y otro incorrecto?

La respuesta a esta pregunta requiere un análisis pormenorizado ítem a ítem. Se ha elaborado una tabla para los dos modelos logísticos y otra para el modelo de Regresión Ordinal. La Figura 6.1 contiene la tabla de los dos modelos logísticos. Para su correcta interpretación se deben tener en cuenta las siguientes premisas:

- En la parte izquierda se presentan los resultados del modelo logístico con variable respuesta `Improve` y en la derecha el modelo logístico con variable respuesta `Positive`.
- En la parte superior se presentan los ítems en los que hay diferencias en el subtitulado y que son objeto de este objetivo específico. En la parte inferior se muestran los ítems que se usan como control ya que no hay diferencias en ellos entre subtitulados y se analizan en el objetivo correspondiente.
- La columna `Freq` es la frecuencia relativa de las tablas de contingencia que resultan del análisis exploratorio. La columna `Prob` son las probabilidades predichas por cada uno de los modelos.
- Los datos se muestran con un fondo coloreado con una tonalidad más oscura cuando más inesperado sea el resultado obtenido.

En el modelo con variable respuesta `Improve` (ver parte izquierda de Figura 6.1), los alumnos han valorado claramente de forma superior el subtitulado *A* que el *B* en los ítems *Q*05 (la corrección ortográfica y gramatical), *Q*06 (la literalidad), *Q*08 (la asignación de líneas a los personajes en los diálogos), *Q*09 (la descripción de efectos sonoros) y *Q*14 (la separación en líneas diferentes de sintagmas nominales, verbales y preposicionales).

En la parte derecha de la Figura 6.1 se muestran las predicciones del modelo con variable respuesta `Positive` de obtener una respuesta con nivel 4 ó 5. Coincide con el modelo anterior en los ítems peor valorados en el subtitulado *B*. Además, habría que añadir el ítem *Q*07 (la identificación de los personajes), que tiene mayor probabilidad de valoración no positiva.

En la Figura 6.2 se muestran las probabilidades por nivel de respuesta, ítem y nivel de subtitulado correspondiente al modelo de Regresión Ordinal y se comparan con las frecuencias de la tabla de contingencia. Los ítems con errores se presentan en negrita y recuadrados. Se observa que en el subtitulado *A* todas las respuestas a los ítems se concentran en valores positivos (4 ó 5). En el subtitulado *B* se espera que los ítems en los que se han introducido errores tengan peores valoraciones. Esto sucede claramente en *Q*05 y en *Q*09 y también aunque en menor medida en *Q*06, *Q*07, *Q*08 y *Q*14. Estos ítems coinciden con los que se han destacado anteriormente y son, por lo tanto, en los que los estudiantes reconocen más fácilmente diferencias entre subtitulados.

	Improve Model		Positive Model			
			Treat A		Treat B	
Item	Freq	Prob	Freq	Prob	Freq	Prob
Evaluated Items						
Q18	74.7%	79.5%	94.3%	97.4%	34.5%	30.4%
Q02	65.5%	68.6%	95.4%	98.0%	52.9%	56.6%
Q05	81.4%	86.5%	95.4%	97.8%	20.9%	14.3%
Q06	72.1%	76.6%	95.3%	97.9%	40.2%	38.5%
Q07	64.4%	67.1%	89.7%	94.9%	42.5%	41.6%
Q08	69.9%	72.8%	81.0%	88.4%	31.4%	26.4%
Q09	83.7%	88.3%	88.4%	94.0%	19.5%	12.4%
Q10	53.2%	55.3%	90.8%	95.6%	51.9%	53.9%
Q11	57.1%	57.5%	94.2%	97.5%	60.0%	66.0%
Q12	61.4%	63.7%	91.9%	96.3%	50.6%	53.4%
Q14	65.3%	70.8%	82.9%	90.5%	32.5%	27.8%
Control Items						
Q01	49.4%	47.9%	90.8%	95.7%	63.2%	70.1%
Q03	51.2%	51.1%	89.5%	94.9%	60.2%	66.4%
Q04	31.0%	26.4%	92.0%	96.5%	86.2%	92.0%
Q13	39.5%	36.1%	93.1%	97.1%	80.2%	87.4%
Q15	44.9%	42.8%	64.5%	74.4%	45.8%	45.5%
Q16	33.3%	34.0%	56.9%	67.5%	41.9%	37.0%
Q17	43.9%	43.2%	65.7%	77.5%	48.0%	45.3%

Figura 6.1: Predicciones de los modelos de Regresión Logística

> **Objetivo específico**
>
> ¿En qué pautas de subtitulado los estudiantes tienen mayor **dificultad** para reconocer diferencias entre un subtitulado correcto y otro incorrecto?

Los dos modelos logísticos (ver Figura 6.1) coinciden en que los estudiantes tienen dificultad para reconocer diferencias en el subtitulado en los ítems:

- *Q*02 (el número de líneas por subtítulo) con probabilidad predicha de mejorar la valoración de 68.56%.

	Treat A										Treat B									
	Data Frequencies					Model Prob.					Data Frequencies					Model Prob.				
Item	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Q18	**1.1%**	**3.4%**	**1.1%**	**36.8%**	**57.5%**	**0.1%**	**1.0%**	**2.7%**	**30.0%**	**66.1%**	**11.5%**	**34.5%**	**19.5%**	**23.0%**	**11.5%**	**5.6%**	**32.9%**	**30.2%**	**27.2%**	**3.3%**
Q01	1.1%	2.3%	5.7%	34.5%	56.3%	0.1%	1.1%	3.1%	32.5%	63.1%	10.3%	20.7%	5.7%	33.3%	29.9%	1.3%	11.2%	21.6%	52.7%	13.0%
Q02	**0.0%**	**1.1%**	**3.4%**	**33.3%**	**62.1%**	**0.1%**	**0.8%**	**2.2%**	**25.8%**	**71.1%**	**5.7%**	**28.7%**	**12.6%**	**33.3%**	**19.5%**	**2.2%**	**17.2%**	**27.0%**	**45.2%**	**8.2%**
Q03	0.0%	7.0%	3.5%	37.2%	52.3%	0.1%	1.4%	3.8%	36.6%	58.1%	6.0%	24.1%	9.6%	33.7%	26.5%	1.6%	13.1%	23.6%	50.4%	11.1%
Q04	0.0%	2.3%	5.7%	31.0%	60.9%	0.1%	0.9%	2.5%	28.0%	68.6%	2.3%	5.7%	5.7%	39.1%	47.1%	0.3%	2.3%	6.2%	46.5%	44.6%
Q05	**0.0%**	**0.0%**	**4.6%**	**31.0%**	**64.4%**	**0.1%**	**0.7%**	**1.9%**	**23.5%**	**73.8%**	**33.7%**	**36.0%**	**9.3%**	**8.1%**	**12.8%**	**18.8%**	**51.3%**	**18.9%**	**9.3%**	**0.9%**
Q06	**0.0%**	**0.0%**	**4.7%**	**36.0%**	**59.3%**	**0.1%**	**0.8%**	**2.3%**	**27.1%**	**69.7%**	**6.9%**	**33.3%**	**19.5%**	**25.3%**	**14.9%**	**3.9%**	**26.3%**	**30.1%**	**34.1%**	**4.7%**
Q07	**0.0%**	**3.4%**	**6.9%**	**39.1%**	**50.6%**	**0.2%**	**1.5%**	**4.1%**	**38.0%**	**56.3%**	**14.9%**	**33.3%**	**9.2%**	**23.0%**	**19.5%**	**4.4%**	**28.4%**	**30.2%**	**31.9%**	**4.2%**
Q08	**0.0%**	**7.1%**	**11.9%**	**29.8%**	**51.2%**	**0.2%**	**1.8%**	**4.9%**	**41.9%**	**51.2%**	**17.4%**	**40.7%**	**10.5%**	**17.4%**	**14.0%**	**7.6%**	**38.8%**	**29.1%**	**21.7%**	**2.4%**
Q09	**1.2%**	**5.8%**	**4.7%**	**37.2%**	**51.2%**	**0.2%**	**1.5%**	**4.0%**	**37.5%**	**56.9%**	**43.7%**	**27.6%**	**9.2%**	**11.5%**	**8.0%**	**25.6%**	**51.7%**	**14.5%**	**6.5%**	**0.6%**
Q10	**0.0%**	**1.1%**	**8.0%**	**42.5%**	**48.3%**	**0.2%**	**1.4%**	**3.9%**	**37.4%**	**57.0%**	**5.1%**	**29.1%**	**13.9%**	**27.8%**	**24.1%**	**2.3%**	**17.8%**	**27.5%**	**44.4%**	**7.8%**
Q11	**0.0%**	**2.3%**	**3.5%**	**45.3%**	**48.8%**	**0.1%**	**1.4%**	**3.7%**	**36.3%**	**58.3%**	**2.4%**	**14.1%**	**23.5%**	**42.4%**	**17.6%**	**1.4%**	**11.8%**	**22.2%**	**52.0%**	**12.4%**
Q12	**0.0%**	**3.5%**	**4.7%**	**48.8%**	**43.0%**	**0.2%**	**1.8%**	**4.9%**	**41.7%**	**51.4%**	**4.8%**	**27.7%**	**16.9%**	**32.5%**	**18.1%**	**2.5%**	**18.6%**	**27.9%**	**43.4%**	**7.5%**
Q13	0.0%	0.0%	6.9%	32.2%	60.9%	0.1%	0.8%	2.3%	27.0%	69.7%	2.3%	2.3%	15.1%	39.5%	40.7%	0.4%	3.3%	8.3%	51.6%	36.3%
Q14	**1.3%**	**3.9%**	**11.8%**	**35.5%**	**47.4%**	**0.2%**	**1.9%**	**5.0%**	**42.4%**	**50.5%**	**10.4%**	**33.8%**	**23.4%**	**22.1%**	**10.4%**	**5.4%**	**32.2%**	**30.2%**	**27.9%**	**3.4%**
Q15	0.0%	13.2%	22.4%	36.8%	27.6%	0.6%	5.5%	13.0%	54.8%	24.7%	6.9%	20.8%	26.4%	26.4%	19.4%	2.5%	18.5%	27.8%	43.5%	7.5%
Q16	0.0%	3.1%	40.0%	35.4%	21.5%	0.6%	5.7%	13.2%	54.8%	24.3%	3.2%	4.8%	50.0%	25.8%	16.1%	2.0%	15.6%	25.9%	47.2%	9.1%
Q17	0.0%	8.6%	25.7%	35.7%	30.0%	0.4%	4.0%	10.0%	53.6%	31.6%	5.3%	12.0%	34.7%	32.0%	16.0%	2.4%	18.1%	27.6%	44.1%	7.7%

Figura 6.2: Predicciones del modelo de Regresión Logística.

- $Q10$ (la sincronización de las entradas y salidas de los subtítulos) con probabilidad de mejora 55.34%.
- $Q11$ (la velocidad de exposición de los subtítulos) con probabilidad 57.53%.
- $Q12$ (el máximo número de caracteres por línea) con probabilidad 63.7%.

En estos mismos ítems, el modelo ordinal (ver Figura 6.2) predice más respuestas negativas en el subtitulado *B* que en el *A* pero aún así el subtitulado *B* tiene un alto número de respuestas positivas.

Para entender las motivaciones de las valoraciones de los alumnos, se han analizado los comentarios que dejaron [1]. La siguiente es una selección de los comentarios más relevantes en cada ítem:

- $Q02$, **el número de líneas por subtítulo**: En los comentarios al subtitulado *A* hay bastantes que se quejan del número excesivo de líneas.

[1] Estos comentarios se separaron en la fase de preprocesado y no se han utilizado ni consultado hasta la realización de este capítulo.

Subtitulado *A*: “Se pueden hasta 3 pero no es recomendable”, “en ocasiones son 3 innecesariamente”, “frases muy cortas”, “No superaba las dos (creo recordar)”, “Se cumple”.

- *Q*10, **la sincronización de las entradas y salidas de los subtítulos**: En los comentarios al subtitulado *A* hay algunos que dicen que hay falta de sincronización y, por el contrario, en el *B* que estaban bien sincronizados. Hay por tanto una falta de atención de algunos estudiantes para evaluar este aspecto del subtitulado.

  Subtitulado *A*: “Van a destiempo”, “a veces hay retardo del texto sobre el audio”, “No me dado cuenta.”, “Estaban desincronizadas”.

  Subtitulado *B*: “Regular.”, “Sincronizado”, “bastante sincronizados”, “va a la paz texto y lenguaje”, “Estaba bien sincronizado”, “Fallos corregidos”, “sincronizadas con el audio y la imagen”.

- *Q*11, **la velocidad de exposición de los subtítulos**: Los comentarios al subtitulado *B* indican que muchos estudiantes no han tenido en cuenta que el subtítulo debe permanecer al menos tres segundos en la pantalla.

  Subtitulado *B*: “Sincronizado.”, “me ha parecido un tiempo suficiente”, “los pude leer bien”, “Buen tiempo para la lectura”, “Se corresponde con los 12 caracteres por segundo”, “velocidad apropiada”, “Velocidad adequada para una buena lectura”.

- *Q*12, **el máximo número de caracteres por línea**: Los comentarios denotan que en general los alumnos conocen el número máximo de caracteres por línea, pero que no se han detenido a medir cuántos hay realmente.

  Subtitulado *A*: “No sobrepasa los 40 caracteres”.

  Subtitulado *B*: “No pasa de 37”, “Se encuentran entre 12 y 37”.

Se aprecia que los alumnos tienen dificultades para valorar las diferencias en la calidad del subtitulado en estos aspectos principalmente porque, aunque conozcan las normas de subtitulado, no han comprobado que se estén cumpliendo en los vídeos de la actividad.

> Objetivo específico
>
> ¿Son los estudiantes capaces de valorar de forma similar los aspectos del subtitulado que no cambian en los vídeos?

Los ítems en los que no se han introducido errores deberían ser valorados de forma similar por los estudiantes. En los modelos logísticos las probabilidades y frecuencias de estos ítems se muestran en la parte inferior de la tabla de la

Figura 6.1 y en el modelo ordinal son las filas no resaltadas de la tabla de la Figura 6.2. Se comprueba que los ítems $Q04$ (el contraste entre los caracteres y el fondo) y $Q13$ (la legibilidad de la tipografía) se valoran, como se esperaba, de forma positiva y similar en ambos subtitulados. Los ítems $Q01$ (la posición de los subtítulos) y $Q03$ (la disposición del texto respecto a la caja donde se muestran los subtítulos) se valoran positivamente en el subtitulado A, pero en el subtitulado B hay una polarización de las valoraciones habiendo muchas positivas y negativas y pocas neutras. Por último, los ítems $Q15$ (la utilización de puntos suspensivos), $Q16$ (la escritura de los números) y $Q17$ (las incorrecciones en el habla) tienen una valoración comparativamente inferior al resto de ítems en el subtitulado A. Esta valoración es incluso inferior en el subtitulado B. Los estudiantes que han realizado comentarios en estos ítems indican que ninguno de ellos es aplicable a los vídeos. Ante esta circunstancia, los alumnos han consignado distintas valoraciones: algunos han contestado «No sé / No contesto», otros han consignado valoraciones neutrales y, finalmente, otros han optado por valoraciones positivas y negativas.

> Objetivo específico
>
> Efecto secuencia: ¿El orden en el que vieron los vídeos los estudiantes influye en la calidad del subtitulado percibida?

> Objetivo específico
>
> Efecto periodo: ¿La evaluación del subtitulado del segundo vídeo visto está influida por haber evaluado un vídeo previamente?

Estos objetivos se responden de forma conjunta por estar ambos efectos relacionados ya que, como se ha explicado, el efecto secuencia en un estudio cruzado AB/BA es la interacción entre el tratamiento y el periodo.

En el modelo ordinal `Response ~ Treat * Period` (ver Sección 4.2) se constató que tanto el periodo como la secuencia son significativos. No obstante, estos efectos son mucho menos importantes que los debidos al subtitulado. Al introducir como variables explicativas el estudiante y el ítem (ver Sección 4.2) tanto el periodo como la secuencia pasan a ser no significativos. En la Sección 5 se comprobó que estos efectos se producen porque la proporción de respuestas de nivel 5 en el subtitulado A sobre el B es superior en el primer periodo que en el segundo y lo contrario ocurre con las de nivel 4. Se concluye que ni el efecto secuencia ni el efecto periodo son importantes al no tener significación estadística.

6.2 Limitaciones del estudio

Aunque el estudio ha permitido responder a la pregunta de investigación, tiene una serie de limitaciones cuya eliminación permitiría ampliar su ámbito:

- Los datos proceden de un MOOC y la actividad fue voluntaria. Hay que suponer que solo los estudiantes más altamente motivados habrán participado en ella.
- El diseño cruzado no requirió un tiempo de lavado (tiempo entre tratamientos) como es habitual en este tipo de diseños. A pesar de que se ha descartado que el efecto periodo o el efecto secuencia hayan tenido una influencia importante en las respuestas a los test, sería interesante controlar el tiempo entre test y asegurar que los participantes han visto ambos vídeos e incluso que, cuando contestan al test, revisan el vídeo en lugar de fiarse de su memoria.
- En el Análisis Exploratorio (ver Sección 4.1) se constató que algunos estudiantes emplearon muy poco tiempo en responder a la actividad y que algunos estudiantes responden siempre con el mismo nivel de respuesta. Sería interesante realizar el estudio eliminando los test de calidad dudosa, para lo cual se debería contar con una muestra mayor.
- Sería interesante comprobar si las respuestas son diferentes si el estudiante dispone del test antes de ver el vídeo.
- En los comentarios de los alumnos se reflejan la existencia de problemas en el subtitulado *B* del ítem $Q01$ (la posición de los subtítulos) y en subtitulado *A* del ítem $Q03$ (la disposición del texto respecto a la caja donde se muestran los subtítulos). En estos ítems no debería haber deficiencias de subtitulado. Sería conveniente que un experto en subtitulado evaluara si realmente los subtítulos son correctos en estos aspectos o si es que no han sido evaluados adecuadamente por los estudiantes.
- Se ha constatado, a través de los comentarios de los alumnos, que no utilizan criterios homogéneos cuando un ítem no es aplicable a los vídeos. Algunos alumnos contestan «No sé / No contesto», como es esperable, pero otros contestan «Neutral» y otros lo hacen negativa o positivamente. Sería deseable dar una información previa a los alumnos de cómo contestar al test.
- Igualmente se ha constatado que los estudiantes tienden a dar puntuaciones más negativas a los ítems cuando saben que se trata del vídeo incorrectamente subtitulado incluso en aquellos ítems en que los vídeos son idénticos.
- Hay ítems, como los relacionados con la tipografía, la posición de los subtítulos o el contraste, en los que al ser los vídeos idénticos, no es posible saber si los estudiantes son capaces de reconocer diferencias en la calidad del subtitulado.
- Sería interesante comparar las respuestas de los estudiantes con las realizadas por un grupo de expertos.

Capítulo 7

Conclusión y trabajo futuro

Este trabajo ha pretendido responder a la pregunta de investigación de si los estudiantes de un curso de accesibilidad son capaces de identificar los errores en el subtitulado de un vídeo, y como objetivos específicos averiguar qué aspectos del subtitulado han sido más fácilmente reconocidos por los estudiantes y en cuáles han tenido más dificultad. Para ello se ha partido de una Exploración Inicial (ver Sección 4.1) y se han propuesto varios modelos estadísticos que tengan en cuenta la naturaleza ordinal y dependiente de la variable respuesta (ver Sección 4.2). Como variables explicativas se ha considerado el nivel de subtitulado, el periodo en el que se ha realizado cada test, la secuencia u orden de realización de los test, el estudiante que ha realizado el test y el ítem al que se responde.

La conclusión más importante es que todos los análisis estadísticos realizados muestran que el nivel de subtitulado es la variable que mejor explica la respuesta a cada ítem. Los efectos secuencia y periodo son comparativamente de poca importancia y se traducen en que en general los estudiantes valoran peor el vídeo visto en el segundo periodo para un mismo nivel de subtitulado. Las variables estudiante e ítem se han tratado como efectos aleatorios por haber considerado que sus observaciones no son independientes. La varianza explicada por la variable estudiante ha sido más grande que la explicada por la variable ítem.

Los estudiantes han identificado errores de subtitulado de los ítems $Q05$ (la corrección ortográfica y gramatical), $Q06$ (la literalidad), $Q07$ (la identificación de los personajes), $Q08$ (la asignación de líneas a los personajes en los diálogos), $Q09$ (la descripción de efectos sonoros) y $Q14$ (la separación en líneas diferentes de sintagmas nominales, verbales y preposicionales).

Sin embargo, han tenido más dificultades en identificar los errores introducidos en los ítems $Q02$ (el número de líneas por subtítulo), $Q10$ (la sincronización de las entradas y salidas de los subtítulos), $Q11$ (la velocidad de exposición de

los subtítulos) y $Q12$ (el máximo número de caracteres por línea). El análisis realizado sobre los comentarios a estos ítems, evidencia que los estudiantes han aprendido las normas que debe regir el subtitulado en estos aspectos pero que no han evaluado minuciosamente si se cumplen realmente. Hay que tener en consideración que esta fue una actividad voluntaria sin incidencia en la calificación del curso, y que en una situación real esto probablemente no habría sucedido ya que habrían realizado una comprobación minuciosa.

En definitiva, los estudiantes conocen las normas de subtitulado y son capaces identificar los errores que no requieren una comprobación exhaustiva. En los que sí lo requieren, que son aquellos que tienen que ver con parámetros temporales (velocidad y sincronización) y espaciales (número de líneas y caracteres por línea), han tenido más dificultad. Esto está en consonancia con Khafik y Pratama (2022). En su estudio analizan el subtitulado en inglés producido por estudiantes cuya lengua nativa no es el inglés y concluyen que los errores más frecuentes son precisamente los que tienen que ver con parámetros temporales y espaciales.

En cuanto a los ítems que tratan aspectos en los que no hay errores en ninguno de los vídeos, los estudiantes han valorado positivamente ambos subtitulados en los ítems $Q04$ (el contraste entre los caracteres y el fondo) y $Q13$ (la legibilidad de la tipografía). Esto no implica necesariamente que, de haber habido deficiencias en estos aspectos, las hubieran reconocido. De hecho, hay evidencias de que los estudiantes noveles tienen dificultades en la identificación de deficiencias en el contraste (ver Molanes-López et al. 2021).

Los ítems $Q15$ (la utilización de puntos suspensivos), $Q16$ (la escritura de los números) y $Q17$ (las incorrecciones en el habla) preguntan cuestiones que no se producen en los vídeos y que son relativamente fáciles de verificar. En las respuestas de los estudiantes han confluido varios problemas verificables a través de los comentarios a los ítems. Por un lado, algunos estudiantes manifiestan que no recuerdan con seguridad la existencia de lo preguntado (puntos suspensivos, números, …). Este problema no es importante ya que es de suponer que en una situación de evaluación de subtitulado real realizarían un segundo visionado de los vídeos para asegurarse. Más preocupante resulta el segundo de los problemas detectados ya que podría estar también presente en otros ítems y haber pasado inadvertido en este trabajo. El problema en cuestión es que en estos ítems la respuesta esperable es «No sé / No contesto». En la Tabla 4.6 se constató que estos son los ítems que más respuestas de este nivel reciben pero que esta respuesta no es masiva. Muchos estudiantes se decantan por valoraciones negativas, positivas o neutrales a pesar de haber indicado en los comentarios que la pregunta realizada no tiene aplicación en la actividad. Este problema es una preocupación general en análisis de escalas de Likert. Por ejemplo, ver Tutz (2020) para una propuesta de modelado estadístico de la categoría neutral en una escala de Likert. Un tercer problema detectado en estos ítems, que es probable que también haya tenido incidencia en otros ítems, es que, a pesar de que los comentarios revelan que los estudiantes piensan que estos ítems no tienen aplicación en ninguno de los vídeos, obtienen peor valoración en el subtitulado B que en el A. Esto estaría indicando que las contestaciones de los estudiantes

sufren cierto «efecto de ventana rota». La hipótesis que aquí se plantea para explicar por qué el subtitulado *B* ha obtenido peores respuestas que el *A* incluso en ítems en los que el estudiante sabe que los subtitulados son idénticos es la siguiente: Hay ítems como *Q*05 (la corrección ortográfica y gramatical) que son fáciles de evaluar y responder por los estudiantes. Si el estudiante encuentra una falta de ortografía en un subtitulado, estaría psicológicamente condicionado a ser más crítico con cualquier otro aspecto del subtitulado. Ante una pregunta que el estudiante no recuerda haber encontrado (por ejemplo, la presencia de puntos suspensivos) tiende a otorgar una valoración inferior en el subtitulado con faltas de ortografía porque considera que existe la posibilidad de haber pasado por alto la presencia de puntos suspensivos. Esto no sucede en todos los casos. Por ejemplo, en la pregunta sobre el contraste (ítem *Q*04), la diferencia entre subtitulados aunque existe es menor. La hipótesis expuesta es coherente con este hecho ya que, mientras que los puntos suspensivos son algo puntual cuya existencia el estudiante sabe que puede pasar inadvertida, el contraste es algo que afecta o puede afectar a todo el subtitulado del vídeo.

Estas conclusiones abren varias vías de investigación que se enumeran aquí a modo de propuesta y sin ánimo de exhaustividad:

- Incorporar al modelo variables como el sexo, la edad, el lugar de nacimiento, el nivel de estudios o el grado de conocimientos de accesibilidad previo. En el momento de realizar este trabajo se disponía de esta información aunque de forma muy incompleta ya que la mayoría de los estudiantes participantes no suministraron la información personal.
- Volver a realizar el análisis completo con los mismos datos pero incorporando desde el principio el conocimiento del subtitulado correcto y siendo más crítico con la calidad de los datos, lo que llevaría a eliminar alguno de los cuestionarios.
- Analizar los datos de la edición del curso de 2023 para ver en qué medida los modelos y las conclusiones se mantienen o cambian.
- Plantear mejoras en la recogida de datos como, por ejemplo, indicaciones detalladas de como responder en caso de duda, desconocimiento, inaplicabilidad, ….
- Sería interesante ver como cambian las respuestas de los estudiantes si en ambos vídeos se introducen errores en el subtitulado en diferentes aspectos.
- Añadir errores de subtitulado para todos o casi todos los ítems.
- Además de las respuestas de los estudiantes, se podría plantear la actividad a profesionales del subtitulado para evaluar las semejanzas y diferencias entre grupos.

Referencias

AENOR (2012). *UNE 153010 Subtitulado para personas sordas y personas con discapacidad auditiva*. Asociación Española de Normalización y Certificación (vid. págs. 1, 22).

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. DOI: 10.1002/9780470594001 (vid. pág. 36).

– (oct. de 2018). *An introduction to categorical data analysis, 3rd Edition*. URL: https://www.wiley.com/en-us/An+Introduction+to+Categorical+Data+Analysis%2C+3rd+Edition-p-9781119405283 (vid. págs. 10, 11).

Barreda S., S. N. (2023). *Bayesian Multilevel Models for Repeated Measures Data: A Conceptual and Practical Introduction in R*. 1st. DOI: 10.4324/9781003285878 (vid. pág. 18).

Bates, D., M. Mächler, B. Bolker y S. Walker (2015). «Fitting Linear Mixed-Effects Models Using lme4». En: *Journal of Statistical Software* 67.1, págs. 1-48. DOI: 10.18637/jss.v067.i01 (vid. pág. 59).

Brant, R. (1990). «Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression». En: *Biometrics* 46.4, págs. 1171-1178. URL: http://www.jstor.org/stable/2532457 (visitado 20-05-2023) (vid. pág. 38).

Bruin, J. (2011). *How do I interpret the coefficients in an ordinal logistic regression in R*. URL: https://stats.oarc.ucla.edu/r/faq/ologit-coefficients (vid. pág. 40).

Bürkner, P.-C. (nov. de 2021a). «Bayesian Item Response Modeling in R with brms and Stan». En: *Journal of Statistical Software* 100. DOI: 10.18637/jss.v100.i05 (vid. pág. 46).

– (2021b). «Bayesian Item Response Modeling in R with brms and Stan». En: *Journal of Statistical Software* 100.5, págs. 1-54. DOI: 10.18637/jss.v100.i05 (vid. pág. 51).

Bürkner, P.-C. y M. Vuorre (feb. de 2019). «Ordinal Regression Models in Psychology: A Tutorial». En: *Advances in Methods and Practices in Psychological Science* 2, pág. 251524591882319. DOI: 10.1177/2515245918823199 (vid. págs. 12, 46, 53).

Chen, D.-G. y J. Chen (ene. de 2021). *Statistical Regression Modeling with R: Longitudinal and Multi-level Modeling*. DOI: 10.1007/978-3-030-67583-7 (vid. págs. 15-17).

Christensen, R. H. B. (2022). *ordinal—Regression Models for Ordinal Data*. R package version 2022.11-16. https://CRAN.R-project.org/package=ordinal (vid. págs. 38-40, 47, 49).

Christensen, R. H. B. (2018). «Cumulative Link Models for Ordinal Regression with the R Package ordinal». En: (vid. pág. 40).

Friendly, M., D. Meyer y A. Zeileis (dic. de 2015). *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data*, págs. 1-525. DOI: 10.1201/b19022 (vid. pág. 11).

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari y D. Rubin (nov. de 2013). *Bayesian Data Analysis*. DOI: 10.1201/b16018 (vid. págs. 16, 19).

Harrell, F. (2020). «Violation of Proportional Odds is Not Fatal». En: URL: https://www.fharrell.com/post/po/ (vid. pág. 14).

Harrell, F. (ene. de 2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. DOI: 10.1007/978-3-319-19425-7 (vid. págs. 38, 40).

Khafik, M. e I. D. Pratama (mayo de 2022). «Analyzing Errors in Students' Subtitle Products». En: *ELE Reviews: English Language Education Reviews* 2.1, págs. 59-73. DOI: 10.22515/elereviews.v2i1.5180. URL: https://ejournal.uinsaid.ac.id/index.php/ele-reviews/article/view/5180 (vid. pág. 74).

Lawson, J. (2015). *Design and Analysis of Experiments with R (1st ed.)* Ed. por Chapman y Hall/CRC. DOI: 10.1201/b17883. URL: https://www.taylorfrancis.com/books/mono/10.1201/b17883/design-analysis-experiments-john-lawson (vid. pág. 7).

Liddell, T. M. y J. K. Kruschke (2018). «Analyzing ordinal data with metric models: What could possibly go wrong?» En: *Journal of Experimental Social Psychology* 79, págs. 328-348. DOI: 10.1016/j.jesp.2018.08.009. URL: https://www.sciencedirect.com/science/article/pii/S0022103117307746 (vid. pág. 9).

Lim, C.-Y. y J. In (jul. de 2021). «Considerations for crossover design in clinical study». En: *Korean Journal of Anesthesiology* 74. DOI: 10.4097/kja.21165 (vid. pág. 7).

Liu, X. (abr. de 2022). *Categorical Data Analysis and Multilevel Modeling Using R*. Ed. por S. P. Ltd. (vid. págs. 14, 15).

Lüdecke, D., M. S. Ben-Shachar, I. Patil, P. Waggoner y D. Makowski (2021). «performance: An R Package for Assessment, Comparison and Testing of Statistical Models». En: *Journal of Open Source Software* 6.60, pág. 3139. DOI: 10.21105/joss.03139 (vid. pág. 16).

Lui, K.-J. (ago. de 2016). *Crossover Designs: Testing, Estimation, and Sample Size*. DOI: 10.1002/9781119114710 (vid. pág. 8).

Molanes-López, E. M., A. Rodriguez-Ascaso, E. Letón y J. Pérez-Martín (2021). «Assessment of Video Accessibility by Students of a MOOC on Digital Materials for All». En: *IEEE Access* 9, págs. 72357-72367. DOI: 10.1109/ACCESS.2021.3079199 (vid. págs. 2, 74).

Nicenboim Bruno Schad Daniel, V. S. (2023). *An Introduction to Bayesian Data Analysis for Cognitive Science*. URL: https://vasishth.github.io/bayescogsci/book/ (vid. pág. 17).

Parton, B. S. (2016). «Video Captions for Online Courses: Do YouTube's Auto-generated Captions Meet Deaf Students' Needs?» En: *Journal of Open, Flexible and Distance Learning* 20, págs. 8-18 (vid. pág. 1).

Pérez Martín, J., A. Rodríguez-Ascaso y E. Molanes-López (nov. de 2021). «Quality of the captions produced by students of an accessibility MOOC using a semi-automatic tool». En: *Universal Access in the Information Society* 20. DOI: 10.1007/s10209-020-00740-9 (vid. pág. 1).

Senn, S. (2022). *Cross-over Trials In Clinical Research*. Ed. por L. John Wiley. DOI: 10.1002/0470854596 (vid. pág. 8).

Takagi, H., S. Kawanaka, M. Kobayashi, T. Itoh y C. Asakawa (2008). «Social Accessibility: Achieving Accessibility through Collaborative Metadata Authoring». En: *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility*. Assets '08. Halifax, Nova Scotia, Canada: Association for Computing Machinery, págs. 193-200. DOI: 10.1145/1414471.1414507. URL: https://doi.org/10.1145/1414471.1414507 (vid. pág. 2).

Tutz, G. (jul. de 2020). «Hierarchical Models for the Analysis of Likert Scales in Regression and Item Response Analysis: Hierarchical Models». En: *International Statistical Review* 89. DOI: 10.1111/insr.12396 (vid. pág. 74).

Uebersax, J. S. (2006). «Likert scales: dispelling the confusion». En: *Statistical Methods for Rater Agreement website*. URL: https://www.john-uebersax.com/stat/likert.htm (vid. pág. 22).

Venables, W. N. y B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. URL: https://www.stats.ox.ac.uk/pub/MASS4/ (vid. pág. 40).

W3C (2018). *Web Content Accessibility Guidelines (WCAG) 2.1*. URL: https://www.w3.org/TR/WCAG21/ (vid. pág. 1).

Yee, T. W. (2023). *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.1-8. URL: https://CRAN.R-project.org/package=VGAM (vid. págs. 38, 40).

Zeileis, A., D. Meyer y K. Hornik (2007). «vcd: Residual-based Shadings for Visualizing (Conditional) Independence». En: *Journal of Computational and Graphical Statistics* 16.3, págs. 507-525. DOI: 10.1198/106186007X237856 (vid. pág. 37).

EFECTO SECUENCIA E INTERACCIÓN TRATAMIENTO VS. PERIODO

Se va a demostrar que el efecto secuencia es equivalente a la interacción de los factores tratamiento y periodo.

A.1 Preparación

Partiendo del siguiente conjunto de datos generado aleatoriamente [1]:

```
set.seed(100)
n <- 1000
df <- data.frame(
    Response = rnorm(n),
    Treat = as.factor(sample(c("A", "B"), n, replace = TRUE)),
    Period = as.factor(sample(c(1, 2), n, replace = TRUE))
)

df$Seq <- as.factor(
    ifelse(
        df$Period == 1 & df$Treat == "A" | df$Period == 2 & df$Treat == "B",
        "AB",
        "BA"
    )
)

head(df, 10)
```

```
     Response Treat Period Seq
1  -0.50219235     B      2  AB
```

[1] Obsérvese que la variable `Response` en esta simulación es cuantitativa y no ordinal. Se ha realizado de esta forma para poder usar un ajuste de mínimos cuadrados en lugar de una regresión ordinal para facilitar el cálculo y su interpretación.

```
2   0.13153117     A      1  AB
3  -0.07891709     A      2  BA
4   0.88678481     A      2  BA
5   0.11697127     A      1  AB
6   0.31863009     A      2  BA
7  -0.58179068     A      2  BA
8   0.71453271     A      1  AB
9  -0.82525943     B      2  AB
10 -0.35986213     B      1  BA
```

Se calculan las medias por cada nivel de factor y combinaciones de niveles que luego serán utilizadas en la interpretación de los coeficientes de los modelos:

```
M <- mean(df$Response) # 1 media de respuesta global

# 2 medias de respuesta para tratamientos A y B
mTreat <- with(df, tapply(Response, Treat, mean))

# 2 medias de respuesta para periodos 1 y 2
mPeriod <- with(df, tapply(Response, Period, mean))

# 2 medias de respuesta para secuencias AB y BA
mSeq <- with(df, tapply(Response, Seq, mean))

# 4 medias de respuesta para las cuatro combinaciones de tratamiento y periodo
m2 <- with(df, tapply(Response, list(Treat, Period), mean))

dTreat <- diff(mTreat) # diferencia de medias entre tratamientos A y B

dPeriod <- diff(mPeriod) # diferencia de medias entre periodos 1 y 2

d2 <- diff(m2) # diferencias entre niveles de tratamiento en cada nivel de periodo
```

A.2 Análisis con un solo factor (tratamiento)

```
l1 <- lm(Response ~ Treat, df)
data.frame(t(coef(l1))) %>% gt()
```

Tabla A.1: Ajuste del modelo Response ~ Treat con contrasts treatment.

X.Intercept.	TreatB
0.03624217	-0.03966751

Se comprueba que el intercepto es la media de la respuesta en el nivel de tratamiento *A*:

```
mTreat[1]
```

```
        A
0.03624217
```

y que la pendiente (parámetro *TreatB*) es la diferencia entre las medias tratamientos:

```
dTreat
```

```
         B
-0.03966751
```

Por ello, para conocer el efecto del tratamiento en el nivel *B* hay que sumar intercepto y pendiente:

```
coef(l1)[[1]] + coef(l1)[[2]] - mTreat[[2]]
```

```
[1] 1.214306e-16
```

Esto es así ya que por defecto R utiliza el contraste conocido como `codificación de tratamiento`:

```
contr.treatment(2)
```

```
  2
1 0
2 1
```

La matriz ampliada añadiendo el intercepto siempre tendrá una columna de 1's:

```
model.matrix(~Treat, expand.grid(Treat = c("A", "B")))
```

```
  (Intercept) TreatB
1           1      0
2           1      1
attr(,"assign")
[1] 0 1
attr(,"contrasts")
attr(,"contrasts")$Treat
[1] "contr.treatment"
```

Cada fila representa el nivel del tratamiento (fila 1 nivel *A* y fila 2 nivel *B*) y las columnas representan los parámetros del modelo. Los valores son los niveles de tratamiento (0 ó 1). Para obtener el significado de cada parámetro, se multiplica el valor del contraste por el parámetro. Así:

- En la primera fila se comprueba que el efecto del tratamiento *A* es el intercepto: $A = 1 \cdot Intercept + 0 \cdot TreatB$.
- En la segunda fila permite comprobar que el valor del parámetro $TreatB$ es la diferencia de los niveles de tratamiento. $B = 1 \cdot Intercept + 1 \cdot TreatB \Rightarrow TreatB = B - Intercept$.

Esto quiere decir que existe una variable para codificar el efecto tratamiento, y esta variable tiene el valor 0 para el nivel *A* por ser el de referencia y 1 para el nivel *B*. La pendiente se codifica como la diferencia del efecto de los dos niveles $(B - A)$.

A.3 Análisis con un dos factores (tratamiento y periodo)

```
l2 <- lm(Response ~ Treat * Period, df)
data.frame(t(coef(l2))) %>% gt()
```

Tabla A.2: Ajuste del modelo Response ~ Treat * Period con contrasts treatment.

X.Intercept.	TreatB	Period2	TreatB.Period2
0.04138614	-0.1076137	-0.01125933	0.1343517

Se comprueba que el intercepto es la media del tratamiento *A* en el periodo 1 por ser estos los valores que R usa como referencia [2]:

```
m2["A", "1"]
```

```
[1] 0.04138614
```

El parámetro $TreatB$ es la diferencia de medias entre los tratamientos en el periodo 1:

```
m2["B", "1"] - m2["A", "1"]
```

```
[1] -0.1076137
```

El parámetro $Period2$ es la diferencia de medias entre los periodos en el nivel de tratamiento *A*:

```
m2["A", "2"] - m2["A", "1"]
```

```
[1] -0.01125933
```

Finalmente, $TreatB : Period2$ es la diferencia entre el segundo periodo y el primero del nivel de tratamiento *B* menos la diferencia entre periodos del nivel de tratamiento *A*:

```
m2["B", "2"] - m2["B", "1"] - (m2["A", "2"] - m2["A", "1"])
```

```
[1] 0.1343517
```

La matriz de contraste nos permite razonar por qué esto es así:

[2] R utiliza como valor de referencia el nivel más bajo de factor.

```
model.matrix(~ Treat * Period, expand.grid(Treat = c("A", "B"), Period = c("1", "2")))
```

```
  (Intercept) TreatB Period2 TreatB:Period2
1           1      0       0              0
2           1      1       0              0
3           1      0       1              0
4           1      1       1              1
attr(,"assign")
[1] 0 1 2 3
attr(,"contrasts")
attr(,"contrasts")$Treat
[1] "contr.treatment"

attr(,"contrasts")$Period
[1] "contr.treatment"
```

- La primera fila es el intercepto y corresponde con el tratamiento A y el periodo 1.
- La segunda fila es el efecto del tratamiento B en el periodo 1 y se calcula con la suma del intercepto y el parámetro $TreatB$. Luego $TreatB$ es la diferencia del efecto de los tratamientos en el periodo 1.
- Análogamente con la tercera fila cse concluye que $Period2$ es la deferencia entre periodos para el tratamiento A.
- Finalmente, la cuarta fila, es el tratamiento B en el periodo 2 y, por lo tanto, $Treat2 : Period2$ es la diferencia el nivel B de tratamiento y el periodo 2 y el nivel de tratamiento A en el periodo 1, menos la diferencia de niveles de tratamiento para el periodo 1 y menos la diferencia de periodos para el tratamiento A.

Obsérvese que antes se ha calculado de forma diferente $TreatB : Period2$. Aplicando la fórmula anterior y se comprueba que produce el mismo resultado:

```
m2["B", "2"] - m2["A", "1"] - (m2["B", "1"] - m2["A", "1"]) - (m2["A", "2"] - m2["A", "1"])
```

```
[1] 0.1343517
```

A.4 Factor secuencia

Se incorpora la secuencia como factor para ver si es equivalente a la interacción entre periodo y tratamiento. En caso de serlo los coeficientes del modelo ajustado deberían coincidir. Sin embargo se constata que los modelos l2 (Tabla A.2) y l3 (Tabla A.3) tienen distintos coeficientes.

```
l3 <- lm(Response ~ Treat + Period + Seq, df)
data.frame(t(coef(l3))) %>% gt()
```

Tabla A.3: Ajuste del modelo Response ~ Treat + Period + Seq con contrasts treatment.

X.Intercept.	TreatB	Period2	SeqBA
0.04138614	-0.04043786	0.05591654	-0.06717587

Los coeficientes no coinciden debido a que se está usando el contraste con `codificación de tratamientos`. Pero si se cambia a `codificación de sumas`:

```
options(contrasts = rep("contr.sum", 2))
```

Y se vuelven a ajustar los modelos (que ya usarán el `contraste suma`), se comprueba que ahora tienen los mismos coeficientes y el coeficiente $Seq1$ del modelo que incorpora el efecto secuencia (Tabla A.4) es igual que el coeficiente $Treat1 : Period1$ del modelo que incorpora la interacción entre tratamiento y periodo (Tabla A.5). Obsérvese que los nombres de los coeficientes han cambiado respecto al `contraste de tratamiento`. Esto sucede porque la interpretación de los coeficientes varía como se explica a continuación.

```
l4 <- lm(Response ~ Treat + Period + Seq, df)
data.frame(t(coef(l4))) %>% gt()
```

Tabla A.4: Ajuste del modelo Response ~ Treat + Period + Seq con contrasts sum.

X.Intercept.	Treat1	Period1	Seq1
0.01553755	0.02021893	-0.02795827	0.03358794

```
l5 <- lm(Response ~ Treat * Period, df)
data.frame(t(coef(l5))) %>% gt()
```

Tabla A.5: Ajuste del modelo Response ~ Treat * Period con contrasts sum.

X.Intercept.	Treat1	Period1	Treat1.Period1
0.01553755	0.02021893	-0.02795827	0.03358794

La interpretación de los coeficientes es diferente. Para explicarlo, se muestra la matriz de contraste:

```
model.matrix(~ Treat * Period, expand.grid(Treat = c("A", "B"), Period = c("1", "2")))
```

```
  (Intercept) Treat1 Period1 Treat1:Period1
1           1      1       1              1
2           1     -1       1             -1
3           1      1      -1             -1
4           1     -1      -1              1
attr(,"assign")
[1] 0 1 2 3
attr(,"contrasts")
attr(,"contrasts")$Treat
[1] "contr.sum"
```

```
attr(,"contrasts")$Period
[1] "contr.sum"
```

Ahora los niveles son 1 y -1 [3] en vez de 0 y 1 que se utilizan en el contraste de tratamiento. La interpretación es la siguiente:

- El intercepto es la media de la media de cada uno de los niveles de factor. ¿Por qué?. El intercepto es el valor de la variable de respuesta cuando cuando todas las variables explicativas valen 0. Esto sucede en la media de la variable de respuesta ya que cero es el valor que está en la mitad de +1 y -1. Se comprueba que la media global coincide con el intercepto del modelo l4 (Tabla A.4):

```
mean(m2)
```

```
[1] 0.01553755
```

- El coeficiente $Treat1$ es la mitad la diferencia de la media entre niveles de tratamiento ($TreatA - TreatB$). La media de cada tratamiento se calcula como la media del tratamiento en cada periodo.

```
-diff(apply(m2, 1, mean)) / 2
```

```
        B
0.02021893
```

Otra forma de entender el coeficiente $Treat1$ es como la cuarta parte de la diferencia de los efectos de los tratamientos en cada periodo.

```
(m2["A", "1"] + m2["A", "2"] - (m2["B", "1"] + m2["B", "2"])) / 4
```

```
[1] 0.02021893
```

- El coeficiente $Period1$ es la mitad la diferencia de la media entre periodos($Period1 - Period2$). La media entre periodos se calcula como la media del periodo para cada tratamiento.

```
-diff(apply(m2, 2, mean)) / 2
```

[3] El nivel de referencia del factor tendrá valor 1 y el otro -1. Por ejemplo, en la variable `Treat`, A tendrá +1 y B tendrá valor -1.

```
         2
-0.02795827
```

Otra forma de entender el coeficiente $Period1$ es como la cuarta parte de la diferencia de los efectos del periodo en cada tratamiento.

```
(m2["A", "1"] + m2["B", "1"] - (m2["A", "2"] + m2["B", "2"])) / 4
```

```
[1] -0.02795827
```

- El coeficiente $Treat1 : Period1$ es el coeficiente $Treat1$ menos la mitad de la diferencia de la media entre tratamientos para el periodo 2 ($TreatA - TreatB$):

```
-diff(apply(m2, 1, mean)) / 2 + diff(m2[, "2"]) / 2
```

```
         B
0.03358794
```

```
coef(l5)[2] + diff(m2[, "2"]) / 2
```

```
    Treat1
0.03358794
```

El coeficiente $Treat1 : Period1$ también se puede calcular como $Period1$ menos la mitad de la diferencia de la media entre periodos para el para el tratamiento B ($Period1 - Period2$):

```
-diff(apply(m2, 2, mean)) / 2 + diff(m2["B", ]) / 2
```

```
         2
0.03358794
```

```
coef(l5)[3] + diff(m2["B", ]) / 2
```

```
   Period1
0.03358794
```

Un tercera forma de interpretar el coeficiente $Treat1 : Period1$ es como la cuarta parte de la suma de la diferencia cruzada del efecto de cada tratamiento en cada periodo:

```
(m2["A", "1"] - m2["A", "2"] + m2["B", "2"] - m2["B", "1"]) / 4
```

```
[1] 0.03358794
```

O reorganizando los términos de otra forma, sería la cuarta parte de la suma de la diferencia cruzada del efecto de cada periodo en cada tratamiento:

```
(m2["B", "2"] - m2["A", "2"] + m2["A", "1"] - m2["B", "1"]) / 4
```

```
[1] 0.03358794
```

- Se puede obtener el coeficiente $TreatB$ del modelo $l2$ (Tabla A.2) como $-2 \cdot (Treat1 + Treat1 : Period1)$:

```
-2 * (coef(l5)["Treat1"] + coef(l5)["Treat1:Period1"])
```

```
    Treat1
-0.1076137
```

- Análogamente el coeficiente $Period2$ del modelo $l2$ (Tabla A.2) se obtiene $-2 \cdot (Period1 + Treat1 : Period1)$:

```
-2 * (coef(l5)["Period1"] + coef(l5)["Treat1:Period1"])
```

```
    Period1
-0.01125933
```

- El coeficiente $TreatB : Period2$ se obtiene como $4 \cdot Treat1 : Period1$:

```
4 * (coef(l5)["Treat1:Period1"])
```

```
Treat1:Period1
     0.1343517
```

A.5 Resumen de modelos y equivalencias de parámetros

En la Tabla A.6 se muestran las equivalencias de los coeficientes de cada modelo. Todas las filas de la misma columna corresponden a un determinado nivel para cada factor y la fórmula mostrada es el modelo resultante teniendo en cuenta que:

- En el contraste `treatment` se utiliza como referencia el primer nivel de cada factor, que corresponderá con el intercepto; los coeficientes se denominan $(Intercept)$, $TreatB$, $Period2$ y $SeqBA$ y son la diferencia del nivel que representa cada coeficiente con el intercepto; y los valores de cada nivel de factor son: $TreatA = 0$, $TreatB = 1$, $Period1 = 0$, $Period2 = 1$, $SeqAB = 0$, $SeqBA = 1$.
- En el contraste `sum`, el intercepto es el valor medio y se excluye elcoeficiente del último nivel que se calcula como la suma del resto de niveles con signo opuesto [4]; los coeficientes se denominan $(Intercept)$, $Treat1$, $Period1$ y $Seq1$; y los valores de cada nivel de factor son:$TreatA = 1$, $TreatB = -1$, $Period1 = 1$, $Period2 = -1$, $SeqAB = 1$, $SeqBA = -1$.

Tabla A.6: Equivalencia entre coeficientes y modelos.

Contrast	Model	Factor Levels			
		TreatA Period1 SeqAB	**TreatA Period2 SeqBA**	**TreatB Period1 SeqBA**	**TreatB Period2 SeqAB**
Treatment	$R = \beta_0 + \beta_1 Treat + \beta_2 Period + \beta_3 Treat:Period$	$R = \beta_0$	$R = \beta_0 + \beta_2$	$R = \beta_0 + \beta_1$	$R = \beta_0 + \beta_1 + \beta_2 + \beta_3$
	$R = \beta_0 + \beta_1 Treat + \beta_2 Period + \beta_3 Seq$	$R = \beta_0$	$R = \beta_0 + \beta_2 + \beta_3$	$R = \beta_0 + \beta_1 + \beta_3$	$R = \beta_0 + \beta_1 + \beta_2$
Sum	$R = \beta_0 + \beta_1 Treat + \beta_2 Period + \beta_3 Treat:Period$	$R = \beta_0 + \beta_1 + \beta_2 + \beta_3$	$R = \beta_0 + \beta_1 - \beta_2 - \beta_3$	$R = \beta_0 - \beta_1 + \beta_2 - \beta_3$	$R = \beta_0 - \beta_1 - \beta_2 + \beta_3$
	$R = \beta_0 + \beta_1 Treat + \beta_2 Period + \beta_3 Seq$	$R = \beta_0 + \beta_1 + \beta_2 + \beta_3$	$R = \beta_0 + \beta_1 - \beta_2 - \beta_3$	$R = \beta_0 - \beta_1 + \beta_2 - \beta_3$	$R = \beta_0 - \beta_1 - \beta_2 + \beta_3$

[4] Como en este caso solo hay dos niveles en cada factor, el valor del segundo nivel será simplemente el opuesto del primer nivel.