
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: 2023-10-03

This document is reproducible thanks to:

- LaTeXand its class memoir (http://www.ctan.org/pkg/memoir).
- R (http://www.r-project.org/) and RStudio (http://www.rstudio.com/)
- bookdown (http://bookdown.org/) and memoiR (https://ericmarcon.github.io/memoiR/)


Name of the owner of the logo

http://www.company.com

# Resumen

TODO: Incluir un resumen del trabajo.

# Agradecimientos

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed malesuada nulla augue, ac facilisis risus pretium a. Ut bibendum risus id ex fermentum, at accumsan erat vulputate. In hac habitasse platea dictumst. Sed lobortis est a enim bibendum, ac pulvinar nulla aliquam. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Pellentesque efficitur justo id suscipit pretium. Proin iaculis sit amet nibh vel euismod. Aenean tincidunt faucibus ex, non vehicula ipsum tristique in. Fusce vel tincidunt lectus, vel rutrum nisi. Suspendisse malesuada lectus ac enim vehicula rhoncus. Nullam convallis justo in bibendum eleifend.

Phasellus vitae magna nec mi sagittis luctus vitae eu augue. Donec scelerisque laoreet arcu, eget tempor mi ultricies vel. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Vestibulum at blandit ex. Vestibulum eu sagittis mauris. In hac habitasse platea dictumst. Duis eget ante vel lacus sollicitudin convallis quis eu velit. Sed auctor sem non nisi hendrerit, vel tincidunt tortor bibendum.

# Índice

# Índice de cuadros

# Índice de figuras

Capítulo 1

# Introducción

Capítulo 2

# Marco teórico y estado del arte

## 2.1 Características del diseño del experimento

El objetivo del estudio es responder a la pregunta de investigación:

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

Este trabajo pretende responder a la pregunta anterior construyendo un modelo estadístico que compare las respuestas de los estudiantes en los dos subtitulados. Como objetivo secundario se analizará si existen diferencias en las respuestas entre diferentes preguntas.

El diseño del experimento es completamente aleatorizado, de respuesta ordinal, cruzado $AB/BA$ y doble ciego. Es decir, que la asignación de los estudiantes a cada grupo fue aleatoria; cada grupo vio los vídeos en orden inverso; los estudiantes no conocían a priori qué vídeo estaban viendo en cada momento y tampoco se disponía de esta información en el momento de realizar el modelado estadístico.

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 del tratamiento es cero utilizando una prueba de aleatorización».

En este trabajo se estudiarán las diferencias existentes entre los dos niveles de subtitulado a través de las respuestas de los alumnos a los ítems de las escalas de Likert de cada uno de los vídeos. Para ello se propondrán modelos estadísticos adecuados al diseño del experimento. La primera cuestión que debemos abordar es que el diseño sea cruzado. Siguiendo a Senn (2022), 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. 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 `carry-over`. Además, el análisis estadístico es más complicado y particularmente cuando la respuesta es ordinal y hay más de dos tratamientos. En la misma línea, Lui (2016) afirma que «el objetivo principal de un diseño cruzado es estudiar la diferencia entre tratamientos individuales (en lugar de la diferencia 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».

En un diseño cruzado debemos preocuparnos de la existencia de los efectos periodo y secuencia (o `carry-over`). El **efecto periodo** aplicado al experimento del subtitulado, se producirá si las respuestas del segundo periodo están influidas por haber realizado la primera actividad de subtitulado. El **efecto secuencia** se producirá si las respuestas fueran diferentes cuando se realizan en un orden que cuando se realizan en el otro.

La segunda 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 (error de tipo S).
- También puede malinterpretarse la interacción entre factores.

Otro factor 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 podemos considerar independientes los ítems que componen el test ya que los ítems pretenden medir la misma variable latente: la calidad del subtitulado.

En los siguientes apartados se proponen distintas técnicas que pretenden responder al objetivo del trabajo teniendo en cuenta las peculiaridades del diseño del experimento comentadas en el apartado anterior.

## 2.2 Correlación entre preguntas con el alfa de Cronbach

Normalmente las preguntas de un cuestionario pretenden medir una variable que está oculta o latente. En nuestro caso es la calidad del subtitulado. Las respuestas a estas preguntas relacionadas deben ser consistentes internamente, es decir, las respuestas deben correlacionarse fuerte y positivamente.

Un índice que se utiliza habitualmente para medir la consistencia interna de un cuestionario es el coeficiente `alfa de Cronbach` (ver Schweinberger 2020). Se define de esta forma:

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N} s_i^2}{s^2}\right) \tag{2.1}$$

Donde:

- $\alpha$ es el coeficiente `alfa de Cronbach`.
- $N$ es el número de ítems de la escala de Likert.
- $s_i^2$ es la varianza de la puntuación del item $i$.
- $s^2$ es la varianza total de las puntuaciones de todos los items.

Valores cercanos 1 indican una fuerte correlación en las respuestas y se admite que las preguntas del cuestionario están midiendo la misma variable latente.

## 2.3 Métodos exploratorios

En esta sección se aplicarán técnicas estadísticas que se basan en tablas de contingencia. Una descripción teórica de este tipo de técnicas se puede encontrar en Agresti (2018). Un tratamiento aplicado y basado en gráficos, que será el enfoque que seguiremos en este trabajo, es realizado en Friendly et al. (2015).

### Asociación de variables con la prueba de homogeneidad $X^2$

La prueba de homogeneidad $\chi^2$ (**leton2021**) está enmarcarda en el esquema $Nominal \leftarrow Nominal$ y contrasta la hipótesis $H_0$ de que no hay diferencias entre grupos frente a $H_1$ de que existen diferencias. Dada una tabla de contingencia (ver Tabla 2.1). El valor del estadístico se calcula:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - cij)^2}{cij}$$

y sigue una $\chi^2_{(r-1)(c-1)}$, donde $c_{ij}$ son las frecuencias esperadas bajo $H_0$ que se calculan:

$$cij = \frac{r_i c_j}{n}$$

Cuadro 2.1: Tabla de contingencia

| | X = 1 | X = 2 | X =3 | |
|---|---|---|---|---|
| Y = 1 | $n_{11}$ | $n_{12}$ | $n_{13}$ | $r_1$ |
| Y = 2 | $n_{21}$ | $n_{22}$ | $n_{23}$ | $r_2$ |
| | $c_1$ | $c_2$ | $c_3$ | $n$ |

### Comparación con *Odds Ratio*

Dada una tabla de contingencia 2 $x$ 2 (ver Tabla 2.2), el *odds* se define como el cociente de las probabilidades complementarias y el *odds ratio* como el cociente de *odds*:

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

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

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

El $OR$ es un índice de asociación relativo entre variables dicotómicas. Ver en **leton2021** el procedimiento de contraste de hipótesis con $OR$.

Cuadro 2.2: Tabla de contingencia de variables dicotómicas

| | X = 1 | X = 0 |
|---|---|---|
| Y = 1 | $n_{11}$ | $n_{12}$ |
| Y = 0 | $n_{21}$ | $n_{22}$ |

## 2.4 Modelos lineales generalizados

Los modelos lineales generalizados ($GLM$) son modelos en los 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: será una distribución de probabilidad de la familia exponencial que se asume que sigue la variable respuesta $Y$.
- Un componente lineal de predictores.

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

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

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

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

### Regresión Logística

La regresión logística (ver ibíd., pp. 68-69) es un caso particular de $GLM$ donde la variable respuesta, $Y$, es Benoulli o Binomial. Es decir, que $Y$ toma valores 0 ó 1. En una función de Bernoulli de parámetro $\pi$, $E[Y] = P(Y = 1) = \pi$. Necesitamos una función que mapee los valores que puede tomar el componente lineal de rango $(-\infty, +\infty)$ a los valores que puede tomar $\pi$ en el rango $(0, 1)$. Una función que puede hacer esto es la función $logit$:

$$logit(Y = 1) = log\left[\frac{P(Y = 1)}{1 - P(Y = 1)}\right] = \alpha + \beta_1 x_1 + ... + \beta_p x_p \qquad (2.2)$$

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

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

Para interpretar los coeficientes, se puede reescribir la ecuación (ver Friendly et al. 2015, p. 260):

$$\frac{P(Y=1 \mid (x_1, x_2, ..., x_p))}{1-P(Y=1 \mid (x_1, x_2, ..., x_p))} = e^{\alpha} e^{\beta_1 x_1} e^{\beta_2 x_2} ... e^{\beta_p x_p}$$

Para una unidad de incremento de $x_1$:

$$\frac{P(Y=1 \mid (x_1+1, x_2, ..., x_p))}{1-P(Y=1 \mid (x_1+1, x_2, ..., x_p))} = e^{\alpha} e^{\beta_1 (x_1+1)} e^{\beta_2 x_2} ... e^{\beta_p x_p}$$

Dividiendo la segunda ecuación entre la primera:

$$\frac{\frac{P(Y=1|(x_1+1,x_2,...,x_p))}{1-P(Y=1|(x_1+1,x_2,...,x_p))}}{\frac{P(Y=1|(x_1,x_2,...,x_p))}{1-P(Y=1|(x_1,x_2,...,x_p))}} = \frac{e^{\alpha} e^{\beta_1 (x_1+1)} e^{\beta_2 x_2} ... e^{\beta_p x_p}}{e^{\alpha} e^{\beta_1 x_1} e^{\beta_2 x_2} ... e^{\beta_p x_p}} = e^{\beta_1}$$

De forma análoga, si en la Ecuación 2.2 suponemos que todas las x's son 0:

$$\frac{P(Y=1 \mid (x_1=0, x_2=0, ..., x_p=0))}{1-P(Y=1 \mid (x_1=0, x_2=0, ..., x_p=0))} = e^{\alpha} e^{0} e^{0} ... e^{0} = e^{\alpha}$$

Es decir:

- $\beta_j$ es el *log OR* asociado a una unidad de incremento de $x_j$.
- $\alpha$ es el *log odds* de $Y$ cuando $x_j = 0, \forall j \in 1...p$.

El contraste de hipótesis para los coeficientes $\beta$:

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

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 (LRT):

$$LRT = \Lambda = -2 \log \frac{L(\widehat{reduced})}{L(\widehat{full})}$$
$$= -2 \log L(\widehat{reduced}) + 2 \log L(\widehat{full}) \sim \chi^2_r \text{ (donde r es el número de } \beta's = 0 \text{ )}$$

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:

$$\begin{aligned} AIC &: -2\log L + 2p \\ BIC &: -2\log L + p\log(n) \end{aligned} \tag{2.3}$$

donde $L$ es el valor de maxima verosimilitud y el segundo sumando es una penalización que será mayor cuanto más complejo sea el modelo.

## Regresión Ordinal

Las respuestas a los ítems de una escala de Likert son ordinales. Ninguna de las técnicas anteriormente expuestas tiene en cuenta esta circunstancia. La Regresión Ordinal es una clase de $GLM$ que comparte muchas similitudes con la Regresión Logística (ver Sección 2.4) pero que tiene en consideración que los valores de la variable de respuesta están ordenados. 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 [1]. Según Bürkner y Vuorre (2019) hay tres clases de Regresión Ordinal:

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

La primera es la más habitual y además la más adecuada para el diseño de experimento realizado (ver ibíd., pp. 23-24) [2]. El modelo acumulativo, $CM$, presupone que la variable ordinal observada, $Y$, proviene de la categorización de una variable latente (no observada) continua, $\tilde{Y}$. Hay $K$ umbrales $\tau_k$ que particionan $\tilde{Y}$ en $K+1$ categorías ordenadas observables (ver Figura 2.1). Si asumimos que $\tilde{Y}$ tiene una cierta distribución (por ejemplo, normal) con distribución acumulada $F$, se puede calcular 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 2.1,

$$Pr(Y = 2) = F(\tau_2) - F(\tau_1)$$

[1] 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.

[2] 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.

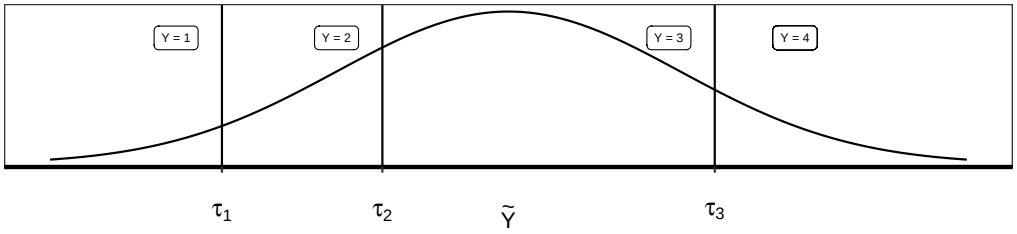



Figura 2.1: Función latente en una regresión ordinal acumulativa.

Si suponemos que $\tilde{Y}$ tiene una relación lineal los predictores:

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

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

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

Y podremos calcular la distribución de probabilidad acumulada de $Y$:

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

Por lo que asumiendo la normalidad de los errores:

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

Donde hay que estimar los umbrales y los coeficientes de regresión. La función anterior es la conocida como la función de enlace `probit`. Otra función de enlace popular es la función `logit`. Es la que usaremos en este trabajo por ser más fácil su interpretación [3]. Con esta función de enlace la interpretación de los coeficientes es parecida a la de los coeficientes de la regresión logística. 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:

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

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

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

Y que [4]:

[3] En la práctica los coeficientes estimados con las funciones de enlace `probit` y `logit` suelen similares.

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

$$\frac{\frac{\Pr(Y\leq k|x_i=1)}{\Pr(Y>k|x_i=1)}}{\frac{\Pr(Y\leq k|x_i=0)}{\Pr(Y>k|x_i=0)}} = \exp(-\beta_i)$$

o, equivalentemente:

$$\frac{\frac{\Pr(Y>k|x_i=x+1)}{\Pr(Y\leq k|x_i=x+1)}}{\frac{\Pr(Y>k|x_i=x)}{\Pr(Y\leq k|x_i=x)}} = \exp(\beta_i)$$

Es decir, que $\exp(\beta_i)$ es el $OR$ (cambio relativo entre *odds*) 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 $x_i$. Un valor del coeficiente $\beta_i$ positivo indica que la relación entre el predictor $x_i$ y la función de *logit* es positiva y, por lo tanto, se incrementa la probabilidad de un mayor valor de la variable respuesta. 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). 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.»

## 2.5 Modelos multinivel

Un modelo multinivel, jerárquico o mixto es un modelo en el que los datos están anidados en una estructura jerárquica. 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. Los modelos multinivel se utilizan cuando se incumple la hipótesis de independencia entre las observaciones. En el caso de los métodos de enseñanza, 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 [5]. 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. Distinguimos 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):

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

donde 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 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 ($\beta_0$ y $\beta_1$), la varianza intra-grupos ($\sigma^2$), la varianza inter-grupos del intercepto aleatoria ($\tau_0$) y de la pendiente aleatoria ($\tau_1$), y la correlación entre intercepto y pendiente aleatorias ($\rho_{01}$).

En Gelman et al. (2013) 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.

[5] 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.

- *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 podemos 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, pp. 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.

## 2.6 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 $p(y \mid \theta)$ que depende del parámetro desconocido $\theta$. En el análisis bayesiano se considera que $\theta$ es una variable aleatoria ya que tenemos incertidumbre respecto a su valor. Esto se traduce en que debemos asignar una distribución de probabilidad $p(\theta)$ conocida como distribución a priori que expresa nuestra creencia sobre los valores que puede tomar $\theta$. En la inferencia bayesiana se usa la distribución de probabilidad a posteriori $p(\theta \mid y)$ 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 que proviene del desconocimiento del valor de $\theta$. En la estimación frecuentista, debido a que se utilizan estimaciones puntuales de $\theta$, no se tiene en cuenta esta incertidumbre. La Ecuación 2.4 se corresponde

con la distribución predictiva posteriori que tiene en consideración ambas incertidumbres:

$$\begin{aligned} p(y_{pred} \mid y) &= \int_{\theta} p(y_{pred}, \theta \mid y)\, d\theta = \int_{\theta} p(y_{pred} \mid \theta, y) p(\theta \mid y)\, d\theta \\ &= \int_{\theta} p(y_{pred} \mid \theta) p(\theta \mid y)\, d\theta \end{aligned} \tag{2.4}$$

donde la última igualdad resulta de la independencia condicional de $y_{pred}$ e $y$ dado $\theta$ ($y_{pred} \perp\!\!\!\perp y \mid \theta$). 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 [6]. 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 Montcarlo basados en Cadenas de Márkov) [7].

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

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

La *lpd* es la densidad conjunta de observar los datos dada la estructura del modelo y las estimaciones de los parámetros $\theta$. Aunque las probabilidades a priori no se incluyen en su cálculo, sí influyen en la estimación de $\theta$ 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

[6] Por eso a estos intervalos se les conoce como intervalos de credibilidad.

[7] En ocasiones se puede obtener una forma analítica 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.

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 podemos saber el valor de *elpd* ya que no conocemos el proceso que genera verdaderos valores $\tilde{y}$. Una forma de estimar *elpd* que empíricamente se ha visto que funciona es penalizar *lpd* con el número de parámetros $p$ de formá análoga a los que se hace en $AIC$ (ver Ecuación 2.3):

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

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 pueden 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. Para evitar tener que dividir el conjunto de datos en datos de entrenamiento y de validación se puede hacer validación cruzada de un solo elemento o $LOO$. En esta validación se deja un elemento fuera cada vez. El problema es que tendremos que estimar el modelo tantas veces como datos tengamos. Para evitar esto, hay formas de aproximar $\widehat{\text{elpd}}$ basadas en $LOO$ sin tener que reentrenar el modelo. Las 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 $\theta$ que resulta tras eliminar la observación $y_i$ [8].

[8] No se entra en detalles de como estimar $\theta_{y_{-i}}$ sin reentrenar el modelo.

Capítulo 3

# Métodología/Materiales y métodos

## 3.1 Descripción de la experiencia

### Marco de la experiencia

Los datos se recogieron de una actividad de subtitulado que se propuso a los estudiantes de la edición 2022 del curso MOOC Materiales digitales accesibles de la UNED Abierta. Este curso pertenece al Canal Fundación ONCE (IEDRA) está dirigido por los profesores Emilio Letón Molina y Alejandro Rodríguez Ascaso y según se recoge en la propia página Web del curso sus objetivos son:

- Reconocer y abordar los desafíos a los que se enfrentan las personas con discapacidad (por ejemplo los estudiantes) al usar los documentos electrónicos, adquirir conciencia y experiencia.
- Obtener una mejor comprensión de la accesibilidad como un asunto de derechos civiles y desarrollar los conocimientos y habilidades necesarios para diseñar recursos de aprendizaje que promuevan ambientes de aprendizaje inclusivos.
- Valorar cómo los documentos accesibles benefician a todas las personas, incluyendo a las que tienen y a las que no tienen discapacidad, a través de una mayor facilidad de uso y la interoperabilidad de los materiales basados en la web.
- Tomar conciencia de que cómo los autores pueden no solo eliminar barreras, sino evitar crearlas en primer lugar.

- Adquirir autosuficiencia en la producción de contenidos accesibles y en la identificación de problemas de accesibilidad.
- Adquirir, como autores, estrategias para la producción sostenible de material digital, como la modularidad.

Está destinado «a todos aquellas personas que desean participar en el desarrollo de soluciones de diseño éticas y creativas, que quieran escribir o gestionar contenidos electrónicos accesibles, desde páginas web o aplicaciones a libros electrónicos»ePub".

El curso tiene cuatro módulos:

- Introducción.
- Accesibilidad de material multimedia.
- Accesibilidad de texto digitales.
- Materiales digitales en la práctica.

## Actividad de subtitulado

La actividad de subtitulado es voluntaria y sin influencia en la calificación final del alumno ni el material al que se tiene acceso. Se realiza en el módulo «Accesibilidad del material multimedia». En este mismo módulo, y antes de proponerles la actividad de subtitulado, los alumnos completaron las secciones «Accesibilidad de la información sonora» y «Accesibilidad de la información visual», con lo cual ya tienen conocimiento sobre creación de vídeos y subtítulos accesibles.

La actividad consistió en ver dos vídeos idénticos y que solo se diferencian en la calidad del subtitulado de 43 segundos de duración. Los subtítulos de uno de los vídeos se realizaron (ver Pérez Martín et al. 2021; Molanes-López et al. 2021) siguiendo la guía Web Content Accessibility Guidelines 2.1 (WCAG 2.1) del W3C (World Wide Web Consortium). 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 es aleatorio, de tal forma que una cohorte 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 items de Likert responden a los criterios de la norma UNE 153010 (ver AENOR 2012). El diseño de experimento es doble ciego: es decir, a los alumnos no se les informó de si estaban viendo el vídeo con mejor o con peor calidad de subtitulado. Esta información tampoco se conoce en el momento de realizar este trabajo ya que los vídeos tienen identificaciones genéricas que no contienen ninguna indicación del tipo de subtitulado del vídeo[1].

[1]En la respuesta a cada ítem, el alumno puede añadir comentarios. Éstos han sido

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

Cuadro 3.1: Niveles de los items 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 |

Cuadro 3.2: Items 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. |

eliminados del estudio para que no filtren información referente al tipo de subtitulado que el alumno cree estar contestando.

[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.

Cuadro 3.3: Tablas de contingencia de la información del sexo, edad y nivel de conocimientos previos.

| gender | Freq |
|---|---|
| femenino | 46 |
| masculino | 19 |
| NA | 22 |

Estudiantes por sexo.

| year_of_birth | Freq |
|---|---|
| None | 22 |
| NA | 1 |

Estudiantes con valor nulo en el campo año de nacimiento.

| level_of_knowledge | Freq |
|---|---|
| 4 | 1 |
| 6 | 2 |
| 7 | 15 |
| 8 | 22 |
| 9 | 20 |
| 10 | 16 |
| NA | 11 |

Estudiantes en función del número de preguntas acertadas en el test de conocimiento sobre accesibilidad.

## 3.2 Participantes

Los datos personales de los estudiantes se suministraron anonimizados para evitar conocer su identidad. De acuerdo con nuestro compromiso ético, del estudio se han eliminado a aquellos 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 sólo realizaron el primero de los test. Como la muestra es suficientemente amplia, se ha decidido eliminar estos test quedando 87 estudiantes participantes. En la Tabla 3.3 se muestran las tablas de contingencia de algunos de los datos que los estudiantes voluntariamente facilitaron en el cuestionario inicial del curso.

## 3.3 Ficheros suministrados

Se dispuso de los siguientes ficheros `csv`:

- El fichero `grade` contiene el identificador de estudiante y el grupo al que pertenece (campo `cohort`).

- El fichero `abo` es la información socioeconómica que voluntariamente ha aportado el estudiante: sexo, año nacimiento, nivel de estudios, ocupación.
- El fichero `conoc` contiene el test de evaluación inicial de conocimientos del estudiante.
- El fichero `exp` es la evaluación del curso realizada por cada estudiante.
- El fichero `acc` contiene la información sobre las necesidades/preferencias de accesibilidad que tiene el estudiante.
- Los 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.

## 3.4 Preprocesado

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

- Se lee el fichero de perfil del usuario. El número de fila con el que el usuario aparece en el fichero se utilizará como identificador del usuario para mantener la trazabilidad y comprobar que las transformaciones realizadas son correctas.

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

- El valor del campo `cohort` se sustituye por una letra, $A$ o $B$, en función del grupo asignado. En el momento de realizar este proceso se desconoce qué vídeo vio primero cada cohorte.

- Se lee el fichero `profile` y se añade información sobre el sexo y el año de nacimiento.

- Se lee el fichero `conoc` y se calcula cuántas preguntas acertó cada usuario en el test de evaluación de conocimientos previos. Se añade esta información al perfil del usuario.

- Se leen los ficheros de test y se procesan. Se utiliza el nombre del fichero (`test1` o `test2`) para saber de qué vídeo se está respondiendo el test [3].

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

- Se seleccionan las preguntas que contienen las respuestas y se renombran para que sea más fácil saber de qué pregunta se trata [4]. Se convierte el campo `LastTry`, que contiene la fecha y hora de realización del test, a formato fecha y hora.

- Se realizan algunas comprobaciones como la ausencia de valores nulos en la variables más relevantes o que no existan inconsistencias ni errores de procesado.

- Se eliminan los comentarios y se graban en fichero aparte para que no revelen información que podría descubrir el tipo de subtitulado que piensa que está evaluando el estudiante.

- Se almacenan los resultados de los test preprocesados en un fichero `csv`.

## 3.5 Variables del modelo.

En la Tabla 3.7 se describen las características más relevantes de las principales variables que se utilizarán en el modelado y en el análisis estadístico.

Cuadro 3.7: Descripción de las variables más importantes.

| Nombre | Descripción | Tipo | Valores |
|---|---|---|---|
| Response | Respuesta a las preguntas del test. | Factor ordenado | De 0 a 5[1] |
| Level | Valoración de la respuesta. | Factor ordenado | Negative, Neutral, Positive[2] |
| Treat | Subtítulos | Factor | A o B[3] |
| Period | Periodo | Factor | 1 ó 2[4] |
| Seq | Secuencia de aplicación de los tratamientos. | Factor | AB o BA |
| Subject | Identificación del estudiante | Factor | Numérico |
| Question | Número de la pregunta | Factor | Q01, Q02, ..., Q18[5] |

[1]Se ha hecho una rotación sobre los valores originales. 0 = No sé, 1 = Muy en desacuerdo, ..., 5 Muy de acuerdo.
[2]Positive cuando Response sea 4 ó 5, Negative cuando sea 1 ó 2 y Neutral para 3.
[3]No se conoce si el tratamiento A es el subtitulado bueno o lo es el B.
[4]1 para el primer vídeo visto y 2 el segundo.
[5]Se ha reorganizado de tal forma que Q18, que es la pregunta resumen, sea el valor primero y de referencia.

Partiendo del `dataframe` que se construyó en el preprocesado (ver Sección 3.4) construimos el `dataframe` que usaremos a partir de este momento. Las operaciones principales que se han realizado han sido:

- Renombrar las variables (ver Tabla 3.7).

[4]En los ficheros suministrados la respuesta a cada pregunta ocupa varios campos. Se selecciona en cada pregunta el que contiene el valor de la respuesta y se convierte a numérico.

- Eliminar del estudio los usuarios que solo han realizado uno de los test.
- Transformar las variables que lo requieran en factores. La pregunta 18 se usará como referencia en el factor `Question`.
- Rotar 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.
- Crear el factor `Level` con los niveles `negative`, `neutral` y `positive` dependiendo de si la respuesta es 1 ó 2, 3, 4 ó 5 respectivamente.
- Transformar el `dataframe` de formato ancho a largo: los ficheros de respuestas se suministran en formato ancho. Es decir, que cada fila es un test que contiene 18 columnas para las respuestas a cada pregunta. Los nombres de las columnas son $Q01$, $Q02$, ..., $Q18$ y tendrán valores de 0 a 5 con las respuestas. La mayoría de los paquetes de R que vamos a usar 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, *Question* y *Response*. En la primera se almacenará el identificador de la pregunta ($Q01$, $Q02$, ..., $Q18$) y en la segunda el valor de la respuesta (de 0 a 5). De esta forma, un test pasará de ocupar una fila y 18 columnas en el formato ancho a 18 filas y dos columnas en el largo.

Se crean dos `dataframes`:

- `df_all` contiene en formato largo todas las respuestas a los test.
- `df_clean` tiene la misma estructura que `df_all` pero en él se han eliminado las respuestas «No sé / No contesto».

`df_all` se utilizará cuando se traten las respuestas como categóricas y, por lo tanto, como no ordenadas. `df_clean` se utilizará cuando se traten las respuestas como ordenadas y por ello no contiene las respuestas con valor «No sé / No contesto».

La estructura de estos `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 ...
 $ Question: 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 "Negative"<"Neutral"<..: 2 2 2 2 2 2 2 2 2..
```

En la Tabla 3.8 se muestran algunos ejemplos datos.

Cuadro 3.8: Muestra del dataframe preparado para el modelado estadístico en formato largo.

| Seq | Period | Treat | Subject | Question | Response | Level |
|---|---|---|---|---|---|---|
| BA | 1 | B | 1217 | Q16 | 5 | Positive |
| AB | 1 | A | 38 | Q01 | 5 | Positive |
| BA | 2 | A | 264 | Q18 | 5 | Positive |
| BA | 2 | A | 320 | Q08 | 2 | Negative |
| AB | 1 | A | 498 | Q15 | 2 | Negative |
| BA | 1 | B | 728 | Q04 | 2 | Negative |
| AB | 2 | B | 1126 | Q18 | 3 | Neutral |
| AB | 2 | B | 624 | Q09 | 2 | Negative |
| AB | 2 | B | 916 | Q10 | 1 | Negative |
| BA | 2 | A | 264 | Q09 | 5 | Positive |

Capítulo 4

# Modelado estadístico

En este capítulo se realizará un análisis inicial para describir los datos y descubrir las relaciones que se han encontrado utilizando técnicas de estadística descriptiva. En la Sección 4.2 se concreta y detalla cómo se aplicarán las técnicas estadísticas propuestas en el Capítulo 2 al diseño de experimento. Se pospone la presentación de resultados al Capítulo 5.

## 4.1 Análisis inicial

Como se explica en la Tabla 3.7, al subtitulado le denominamos 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 3.1). Tras eliminar los test de los usuarios que no dieron su consentimiento para participar en el estudio y los de los que no realizaron el segundo test, las dos cohortes están equilibradas (ver Figura 4.1).

### 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 empleó cada estudiante entre

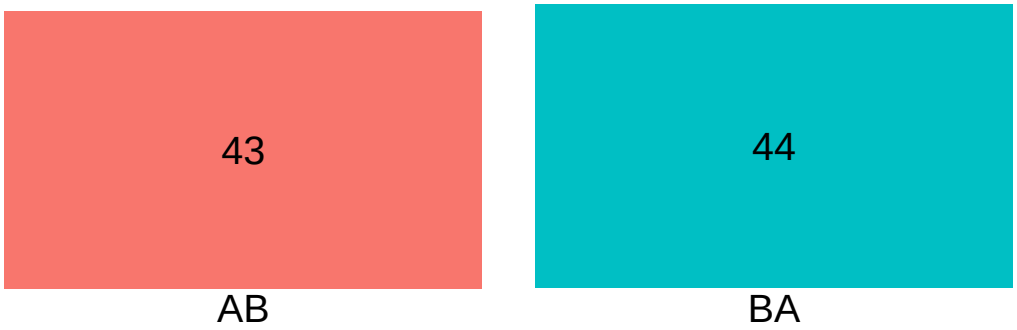



Figura 4.1: Estudiantes asignados a cada grupo.

actividades. La Tabla 4.1 muestra que hay algunos test que se hicieron demasiado rápido [1].

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

| Minutes |
|---|
| 0.93 |
| 1.3 |
| 1.7 |
| 1.72 |
| 1.78 |
| 1.97 |

La Figura 4.2 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 todas los ítems con 1 respuesta.

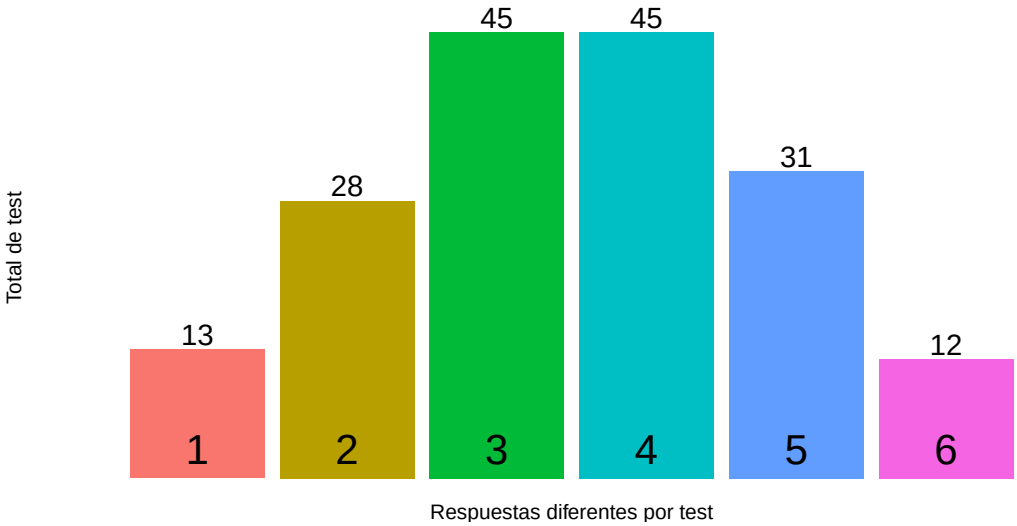



Figura 4.2: Número de respuestas diferentes en un mismo test.

[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.

La tabla Tabla 4.2 muestra los test de respuesta única y el valor de esa respuesta.



Cuadro 4.2: Test en los que todas las preguntas se contestan el mismo valor de respuesta.

| Response | Seq | Test |
|---|---|---|

| | | |
|---|---|---|
| 3 | BA | 02 |
| 3 | BA | 02 |
| 3 | BA | 02 |
| 4 | AB | 01 |
| 4 | AB | 01 |
| 4 | AB | 02 |
| 4 | BA | 01 |
| 4 | BA | 02 |
| 4 | BA | 02 |
| 4 | BA | 02 |

La Figura 4.3 presenta la distribución de la cantidad de respuestas cuyo valor cambia entre los dos test que realiza cada estudiante.

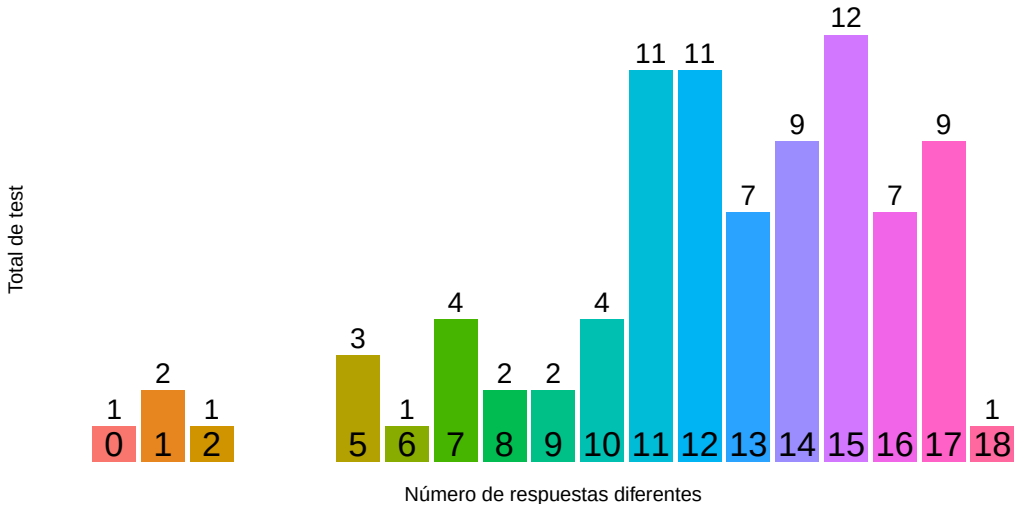



Figura 4.3: Número de respuestas diferentes entre los test para cada estudiante.

Tan solo 1 estudiante respondió a todas las preguntas 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).

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

| Test | Total respuesta por test |
|---|---|
| 01 | 5 |
| 01 | 5 |
| 02 | 5 |
| 02 | 5 |
| 01 | 4 |

Vemos que algunos test tienen valores que no parecen muy razonables. Por ejemplo, no parece razonable realizar la actividad en menos de 2 minutos. Se observa que en algunos test hay poca variabilidad. 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 tratamientos $A$ y $B$ entre grupos.

La Figura 4.4 presenta una forma de comparar los dos test realizados por los estudiantes. Para cada estudiante se comparó pregunta a pregunta sus dos test y se contabilizó la diferencia entre el número de preguntas en que la puntuación en el segundo vídeo fue superior y en las que lo fue inferior (las que no variaron de puntuación no se consideraron). En el eje $x$ se muestra la diferencia entre preguntas. 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 es negativa [2]. Esto es una forma de evaluar si el estudiante valoró mejor o no el segundo vídeo que el primero.

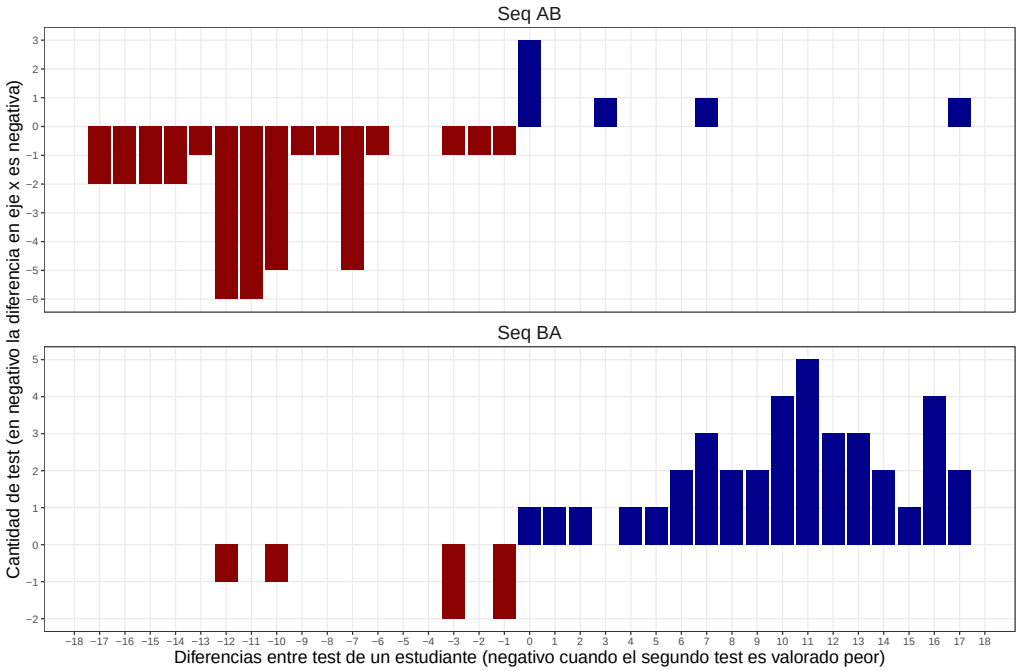



Figura 4.4: Frecuencias absolutas de las diferencias en las respuestas entre test por estudiante y grupo.

Vemos 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

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

valor es 11. f 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.

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

| Seq | Diff | Minutes |
|---|---|---|
| AB | 17 | 1.3 |
| AB | 7 | 3.33 |
| BA | -10 | 50345.95 |
| BA | -12 | 1.7 |

En la Figura 4.5 se muestra la frecuencia relativa del valor de respuesta para cada grupo y test en todas la preguntas [3]. Esta es otra forma de comparar los niveles de subtitulado.

Cuadro 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 |

La Figura 4.5 muestra algunas cuestiones interesantes:

- El tratamiento (subtitulado) con nivel $A$ presenta claramente mayores valores de respuesta que el $B$ como ya habíamos visto (ver Figura 4.4).
- En general los dos grupos 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 frente a un 84% el grupo $BA$. 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.

[3] En el Tabla 4.5 se presenta la misma información con los valores absolutos.


Figura 4.5: Frecuencias relativas de las respuestas al test.

El análisis marginalizado de tratamiento, secuencia y periodo tiene estos resultados referidos a las preguntas 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 las preguntas.

El gráfico Figura 4.6 muestra la frecuencia relativa por grupo y por test de las preguntas clasificadas 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.

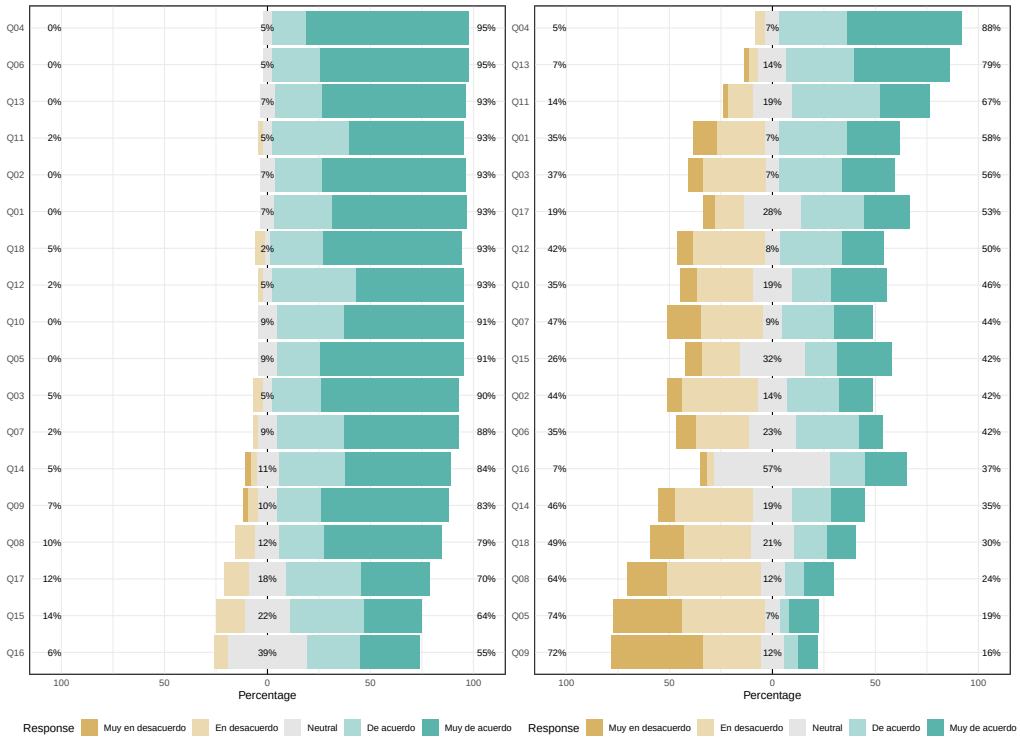
Se muestra en primer lugar la pregunta 18 por ser una valoración global del subtitulado y que resume la opinión que sobre el mismo tiene el estudiante. Volvemos a constatar que el subtitulado $A$ es mejor valorado por los estudiantes, pero ahora vemos que en las 18 preguntas ambos grupos tienen más puntuaciones positivas y menos negativas en el subtitulado $A$ que el $B$. También volvemos a encontrar que los dos grupos valoran de forma muy similar los dos niveles de subtitulado en todas las preguntas. En el nivel de subtitulado $A$ las preguntas $Q15$, $Q16$ y $Q17$ obtienen relativamente peores valoraciones (consultar la Tabla 3.2 para ver el texto de las preguntas) y estas son similares en ambos subtitulados. Hay algunas preguntas que son valoradas de forma positiva incluso en el nivel de subtitulado $B$ (por ejemplo $Q04$ o $Q13$) y que, por lo tanto, su valoración es similar en ambos subtitulados. Por último, las preguntas $Q05$ y $Q09$ (también la $Q14$ pero solo para el grupo $BA$) tienen una valoración muy negativa en el nivel de subtitulado $B$.


Figura 4.6: Frecuencias relativas de las respuestas por pregunta.

La figura Figura 4.7 clasifica las preguntas por valoración y permite constatar lo que ya habíamos visto en el párrafo anterior con mayor comodidad.

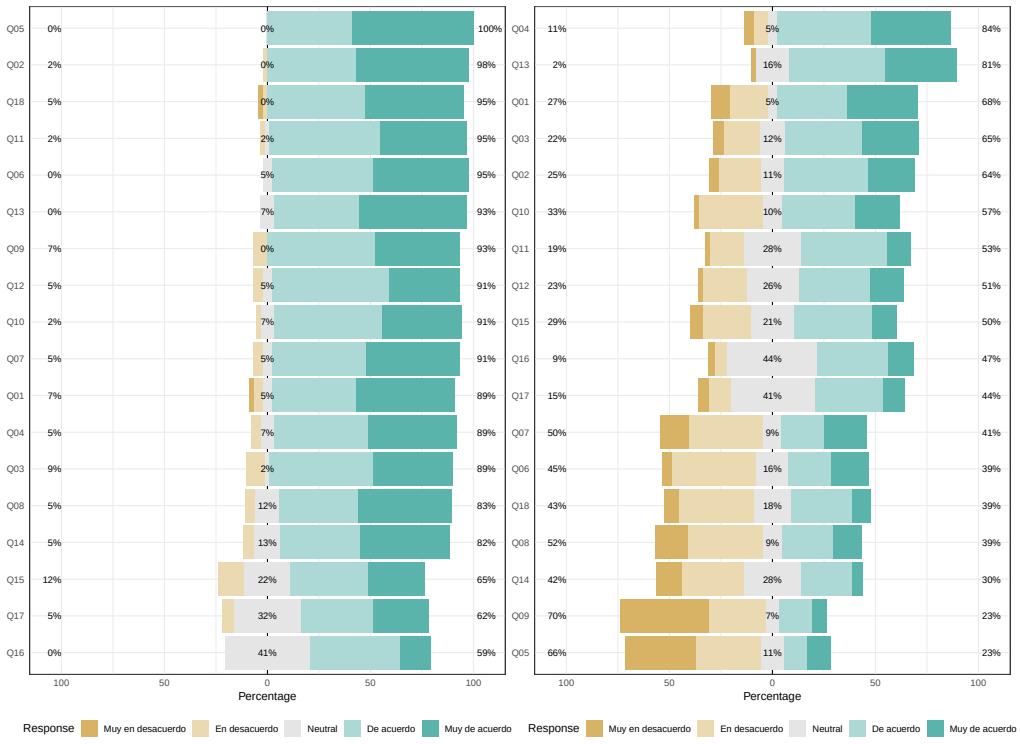
En la Figura 4.8 se muestra para cada pregunta la proporción de estudiantes que han valorado mejor el subtitulado $A$ que el $B$ [4]. Se comprueba que la

[4]Se han eliminado las preguntas en las que una de las dos respuestas del usuario ha




(a) Seq AB , Treat A

(b) Seq AB , Treat B




(c) Seq BA , Treat A

(d) Seq BA , Treat B

Figura 4.7: Preguntas ordenadas por valoración.

mayoría de las preguntas superan el 50%, lo que indica que los estudiantes valoran mejor el subtitulado $A$.

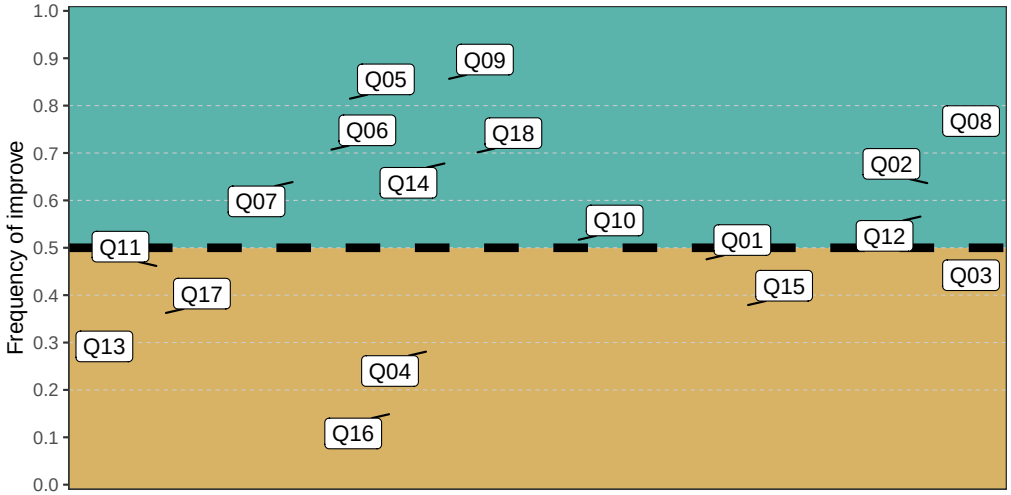



Figura 4.8: Frecuencia de preguntas que mejoran por estudiante entre subtitulados (A > B)

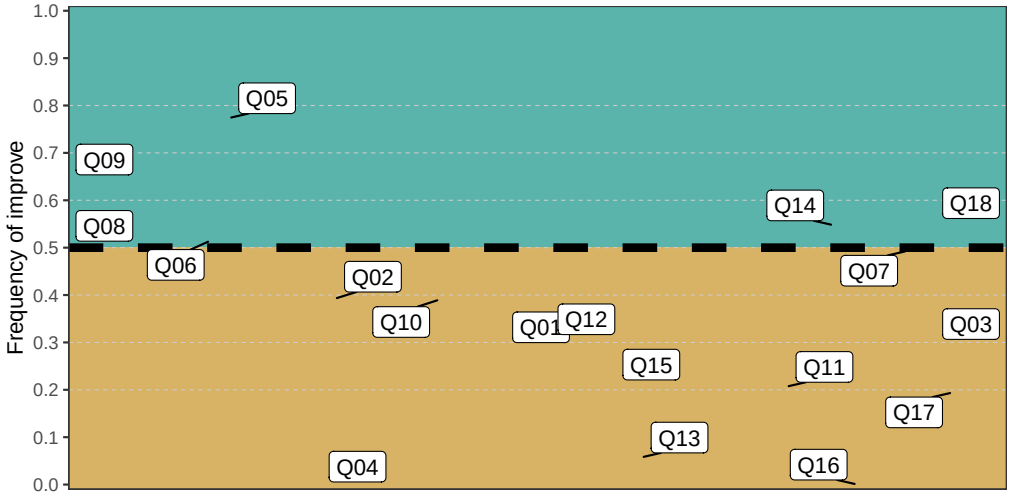



Figura 4.9: Frecuencia de preguntas que mejoran por estudiante entre subtitulados (positive A vs negative B)

En la Figura 4.9 se hace una comparación más exigente ya que ahora se muestra la frecuencia con la que los estudiantes cambian el nivel del subtitulado: de haber valorado el subtitulado positivamente (4 ó 5) en $A$ a hacerlo de forma negativa en $B$ (1 ó 2). En esta ocasión solo las preguntas Q18, Q05, Q06, Q08, Q09, Q14 superan el 50%.

sido «No sé / No contesto». La comparación realizada es respuesta en $A > B$ frente a $A <= B$.

## 4.2 Modelos utilizados

Algunas de las técnicas de análisis y modelado estadístico propuesto en la Sección 4.2 requieren concreción y adaptación para poder ser utilizadas en el diseño de experimento de subtitulado. En esta sección se explica la forma concreta de aplicarlas y los paquetes y configuración en R utilizados. También se justifican las variables incluidas en cada modelo. La presentación de los resultados se pospone al Capítulo 5.

### Comparación con *Odds Ratio*

En la sección Sección 2.3 se explicó el fundamento teórico de esta técnica. Aquí se expone como se puede aplicar al diseño de experimento que se está analizando. Dado que los factores $Treat$, $Period$ y $Seq$ tienen todos 2 niveles, podemos contrastar si hay interacción entre ellos para cada nivel de respuesta. Es decir, se contrasta la hipótesis $H_0 : OR = 1$ de ausencia de interacción frente a $H_1 : OR \neq 1$ de existencia de interacción 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, podemos aceptar que la hipótesis nula de que los grupos responden de forma similar a cada nivel de subtitulado. Se hará un test similar pero entre subtitulado y periodos. Para realizar el contraste de hipótesis se utilizará la función *loddsratio* del paquete *vcd*.

### Regresión Logística

En la Sección 2.4 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 transformado el `dataframe` de tal forma que si un usuario valoró una pregunta mejor en el subtitulado $A$ que en el $B$, se consigna 1 en la variable respuesta, si empeoró o puntuó igual se consigna 0. Si en uno de los test valora una pregunta con «No sé / No contesto», se elimina esa pregunta.

Se ajusta el modelo con la secuencia como predictor. Se constata que el coeficiente del intercepto es positivo y significativo. El intercepto es el *log odds* de mejorar la valoración en $A$ sobre $B$ respecto a empeorar la valoración. Sin embargo, la secuencia no resulta significativa y además añadirla apenas reduce la «deviance», por lo que el modelo nulo sin predictores resulta más parsimonioso. Se podrían añadir como predictores las preguntas y los estudiantes. No se hace aquí y se pospone la sección de resultados.

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

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.336  -1.257   1.027   1.100   1.100

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.36572    0.08563   4.271 1.94e-05 ***
SeqBA       -0.18153    0.11913  -1.524    0.128
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1575.8  on 1151  degrees of freedom
Residual deviance: 1573.5  on 1150  degrees of freedom
AIC: 1577.5

Number of Fisher Scoring iterations: 4
```

## Regresión Ordinal

En la Sección 2.4 se presentó el fundamento teórico de la Regresión Ordinal Acumulativa. En esta sección se ajustan algunos modelos y se interpretan los resultados.

### Ajuste del modelo ordinal `Response ~ Treat`

Existen varios paquetes en R que permiten ajustar una Regresión Ordinal con función de enlace logística. El más popular es el paquete `Ordinal` (Christensen 2022). El paquete `VGAM` (Yee 2023) es más flexible y potente. Otra posibilidad es usar la función `polr` del paquete `MASS` (Venables y Ripley 2002). Finalmente la función `orm` del paquete `rms` también permite hacerlo (ver Harrell 2015). En este trabajo usaremos el paquete `Ordinal` por permitir también incluir efectos aleatorios que utilizaremos en un apartado posterior. Comenzamos con un modelo simple que tiene como único predictor el nivel de subtitulado por ser la variable objetivo de nuestro modelo:

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

```
formula: Response ~ 1 + Treat
```

```
data:    df_clean

 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
```

El método `summary()` muestra la información resumen. Para su interpretación vamos 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 [6]. 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) podemos, por ejemplo, hacer la siguiente interpretación del significado de este coeficiente referido a dos niveles consecutivos de respuesta:

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

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 $OR$:

[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.

[6] El test de Wald es un contraste de hipótesis estadístico en el que se evalúa si el valor estimado es cero suponiendo que $W = \left(\frac{\hat{\theta}-\theta_0}{se(\hat{\theta})}\right)^2 \sim \chi^2$ .

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

Se comprueba que el $OR$ es equivalente en todos los niveles de respuesta al cuestionario. Esta es una de las suposiciones de la regresión ordinal acumulativa. 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 coeficiente $\beta$ para `Treat` es el `log OR` de observar una mejor respuesta en una pregunta 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:

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

En Tabla 4.6 se muestran las probabilidades para ambos niveles de subtitulado y todos los posibles valores de respuesta.

Cuadro 4.6: 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 |

### Ajuste del modelo ordinal `Response ~ Treat * Period`

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

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

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 4.7 se comparan los coeficientes de los cuatro modelos que se listan a continuación:

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

Cuadro 4.7: 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 |

Comprobamos 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 diferente 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 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 menor valor de respuesta.
- La diferencia en `logits` de los efectos totales del subtitulado es el doble de 0.874.
- 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.
- 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 4.8 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.

Cuadro 4.8: Equivalencia entre los coeficientes calculados con `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 contra ste `sum` para constatar que los tres coeficientes son significativos:

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

 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

Al ser los dos modelos anidados, se pueden comparar con la prueba de razón de verosimilitud. Se comprueba que el segundo 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_treat            Response ~ 1 + Treat      logit flexible
clm_sum_treat.period Response ~ Treat * Period logit flexible

                     no.par    AIC  logLik LR.stat df Pr(>Chisq)
clm_treat                 5 7942.2 -3966.1
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
```

### Comprobación de las hipótesis del modelo

La principal hipótesis de un modelo de Regresión Logística Ordinal Acumulativa es que los coeficientes son iguales entre cualesquiera dos niveles de respuestas correlativos. Se han propuesto diversas fórmulas para comprobar esta hipótesis. El paquete `Ordinal` 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. Se constata que el test resulta significativo para `Treat` y para `Treat:Period`, por lo que para estas dos variables no se puede asumir que los coeficientes estimados se mantengan constantes en todos los niveles de respuesta:

```
Tests of nominal effects

formula: Response ~ Treat * Period
             Df  logLik    AIC     LRT Pr(>Chi)
<none>          -3953.0 7920.0
Treat         3 -3904.4 7828.9  97.172   <2e-16 ***
Period        3 -3951.4 7922.7   3.307   0.3467
Treat:Period  9 -3884.8 7801.6 136.408   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lo que procede es ajustar el modelo relajando la constante de proporcionalidad de esas variables. Se ha realizado esto utilizando la función `vglm` del paquete `VGAM`. Vemos que ahora hay cuatro coeficientes para cada una de las variables `Treat` y `Treat:Period`:

```
                          .
(Intercept):1     4.00775235
(Intercept):2     1.90187023
(Intercept):3     0.91341576
(Intercept):4    -0.67031251
Treat1:1          1.91271777
Treat1:2          1.29001334
Treat1:3          0.98941822
Treat1:4          0.69859893
Period1           0.07308677
Treat1:Period1:1 -0.02478844
Treat1:Period1:2 -0.01955805
Treat1:Period1:3 -0.02173532
Treat1:Period1:4  0.24859758
```

El incremento del número de coeficientes complica mucho su interpretación. Algunos autores (ver Sección 2.4) consideran que la Regresión Lineal Acumulativa resulta útil incluso aunque no se cumpla la constante de proporcionalidad.

## Regresión Ordinal Multinivel

En la Sección 2.5 se expuso el fundamento teórico de estos modelos. Aquí se justificará su interés aplicado al caso del subtitulado de vídeos. Hay dos variables susceptibles de ser incorporadas al modelo como efectos aleatorios.

Una primera variable candidata es el factor `Subject`. Es evidente que los estudiantes representan una muestra de una población más amplia que estaría constituida por todos los estudiantes de todos los cursos de accesibilidad. Pero es que además cada estudiante responde a cada ítem dos veces y, por lo tanto, sus observaciones no son independientes. Por otro lado, las preguntas no son independientes unas de otras 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 las otras preguntas. En Bürkner (2021, pp. 14-16) Bürkner y Vuorre (2019, pp. 19-20) podemos 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 uno en el que únicamente se incorpora a los estudiantes con interceptos variables:

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

donde $ij$ es la observación $i$ del estudiante $j$. Obsérvese que ahora los interceptos $\tau_{kj}$ se descomponen en una parte fija y común para todos los estudiantes $\tau_k$ y una parte variable específica para cada estudiante $U_{0j}$. Para ajustar el modelo se va a utilizar la función `clmm()` del paquete `Ordinal` ya que permite la inclusión de efectos aleatorios.

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

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

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

 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) y 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 4.10 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).

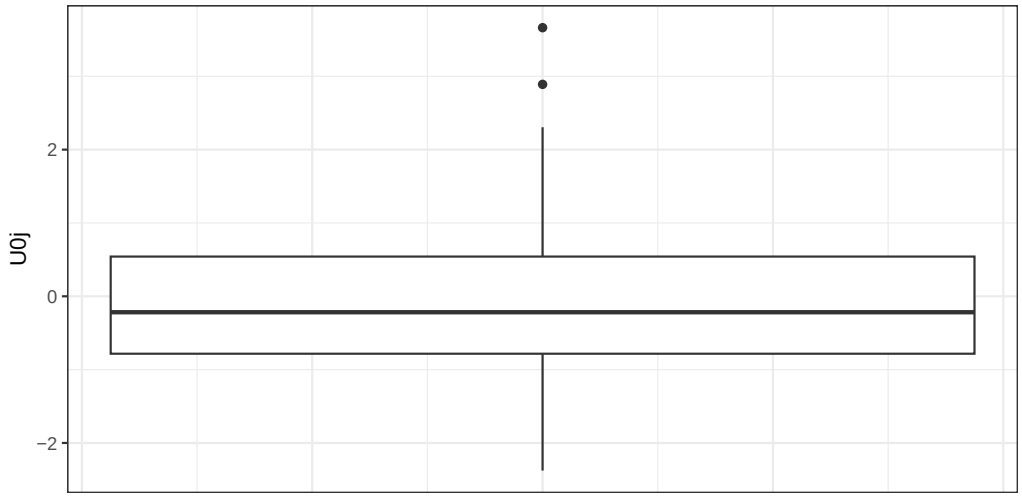



Figura 4.10: Distribución de interceptos aleatorios en el modelo Response ~ Treat * Period + (1 | Subject)

### 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:

$$\begin{aligned}
&\textit{Level } 1 : \text{logit}(P(Response_{ij} \leq k)) = \tau_{kj} - \beta_{1j}\text{Treat}_{ij} - \beta_2\text{Period}_{ij} - \beta_3\text{Treat}_{ij} * \text{Period}_{ij} \\
&\textit{Level } 2 : \tau_{kj} = \tau_k + U_{0j} \\
&\qquad\qquad \beta_{1j} = \beta_1 + U_{1j}
\end{aligned}$$

Ahora el parámetro $\beta_{1j}$ del subtitulado tiene dos componentes: Uno común a todos los estudiantes $\beta_1$ y otro particular de cada estudiante $U_{1j}$. 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.

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

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

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

 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`. Se comprueba que en este test resulta significativamente mejor el último modelo:

```r
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

Se han comparado distintos modelos:

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

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 `Question`. En la Figura 4.11 se muestran las predicciones del modelo. El resumen de parámetros del modelo es el siguiente:

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

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

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

 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
 Question (Intercept) 0.4831   0.6950
          Treat1      0.4655   0.6823   -0.528
Number of groups:  Subject 87,  Question 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
```

Las ecuaciones del modelo están en Ecuación 4.2.

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

donde $ijk$ se corresponde con la observación $i$-ésima del estudiante $j$ y nivel de subtitulado $k$. 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 nivel de subtitulado.

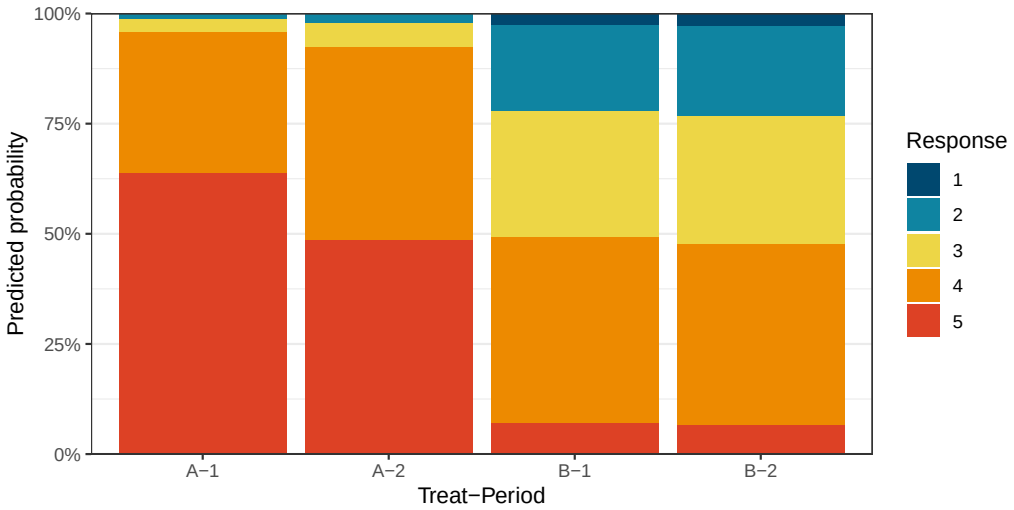



Figura 4.11: Probabilidades de respuesta para el modelo ordinal Response ~ Treat * Period + (1 + Treat | Subject) + (1 + Treat | Question)

## Modelado bayesiano

Existen muchos paquetes en R para hacer inferencia bayesiana. Algunos de ellos son:

- OpenBUGS y WinBUGS: basado en el muestreo de Gibbs.
- JAGS: también utiliza el muestreo de Gibbs.

- Stan: Más moderna y con una comunidad de desarrollo más activa que los anteriores. Utiliza muestreo HMC (Hamiltonian Monte Carlo) y NUTS (no U-turn sampler). Se pueden definir modelos directamente con el lenguaje de modelado de Stan. 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*.
- 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 2.6. 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 | Question)

```
options(contrasts = rep("contr.sum", 2))
brm_treat.period.subject.question <- brm(
    Response ~ Treat * Period + (1 + Treat | Subject) + (1 + Treat | Question),
    data = df_clean,
    family = cumulative("logit"),
    iter = 4000,
    sample_prior = TRUE,
    file = "models/brm_treat.period.subject.question",
    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 las preguntas 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.question)
```

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

Group-Level Effects:
~Question (Number of levels: 18)
                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
sd(Intercept)            0.78      0.15     0.54     1.13 1.00     1873
sd(Treat1)               0.76      0.15     0.53     1.12 1.01     1708
cor(Intercept,Treat1)   -0.46      0.20    -0.79    -0.00 1.00     1738
                     Tail_ESS
sd(Intercept)            3754
sd(Treat1)               3645
cor(Intercept,Treat1)    2863
```

```
~Subject (Number of levels: 87)
                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
sd(Intercept)            1.54      0.14     1.30     1.83 1.00     1339
sd(Treat1)               1.21      0.11     1.02     1.45 1.00     1644
cor(Intercept,Treat1)   -0.11      0.13    -0.36     0.13 1.00     1213
                     Tail_ESS
sd(Intercept)            3047
sd(Treat1)               3421
cor(Intercept,Treat1)    2301

Population-Level Effects:
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept[1]     -4.96      0.27    -5.50    -4.42 1.00      985     1681
Intercept[2]     -2.60      0.26    -3.09    -2.11 1.00      899     1788
Intercept[3]     -1.32      0.25    -1.80    -0.82 1.00      876     1748
Intercept[4]      1.24      0.25     0.76     1.74 1.00      872     1775
Treat1            1.45      0.23     1.01     1.93 1.00     1229     2318
Period1           0.17      0.14    -0.10     0.43 1.00     1004     2236
Treat1:Period1    0.14      0.17    -0.19     0.47 1.01      698     1414

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 4.9 se muestran las distribuciones a priori de los parámetros aleatorios del modelo. En la Figura 4.12 se constata que toman valores razonables y no informativos.

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

| 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 |
| | Intercept | 1 | | | | | | | default |
| | Intercept | 2 | | | | | | | default |
| | Intercept | 3 | | | | | | | default |
| | Intercept | 4 | | | | | | | default |
| lkj_corr_cholesky(1) | L | | | | | | | | default |
| | L | | Question | | | | | | default |
| | L | | Subject | | | | | | default |
| student_t(3, 0, 2.5) | sd | | | | | | 0 | | default |
| | sd | | Question | | | | | | default |
| | sd | Intercept | Question | | | | | | default |
| | sd | Treat1 | Question | | | | | | default |
| | sd | | Subject | | | | | | default |
| | sd | Intercept | Subject | | | | | | default |
| | sd | Treat1 | Subject | | | | | | default |

Es importante asegurar que el entrenamiento ha convergido a su distribución a posteriori. En la tabla de resumen constatamos que el valor de `Rhat` es inferior a 1.1 y el de `ESS` superior a 400 en todos los parámetros, que

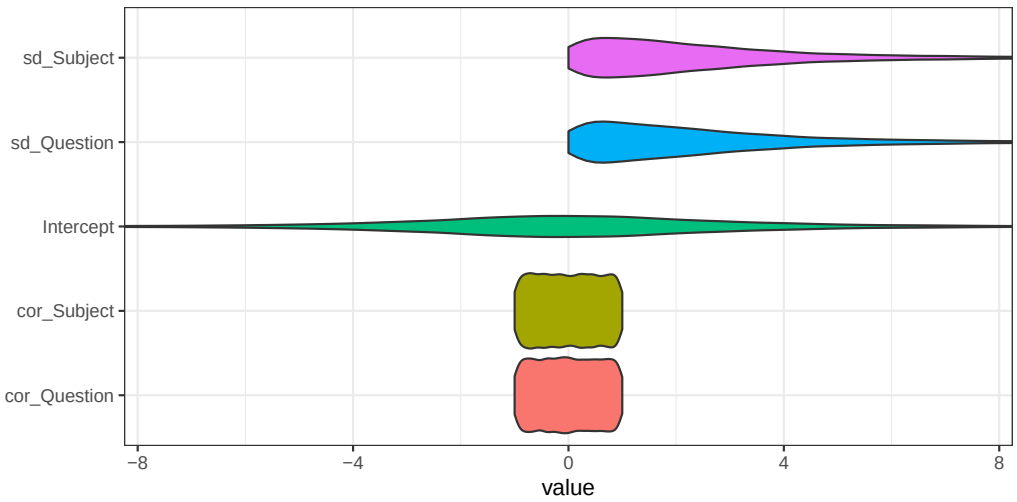



Figura 4.12: Distribuciones a priori del modelo Response ~ Treat * Period + (1 + Treat | Subject) + (1 + Treat | Question).

son umbrales que no se deberían violar (ver Bürkner y Vuorre 2019). En la Figura 4.13 se comprueba que las cadenas MCMC de muestreo de la distribución a posteriori se mezclan correctamente y no se aprecia autocorrelación en ninguno de las parámetros. Por último, en la Figura 4.14 se muestra una comparación entre los histogramas construidos con los datos con los intervalos de confianza marginales de la función predictiva a posteriori del modelo. En la mayoría de las preguntas, el muestreo reproduce bastante bien el histograma de respuestas, aunque en algunas preguntas, como la `Q16` o la `Q17`, se aprecian diferencias relevantes.

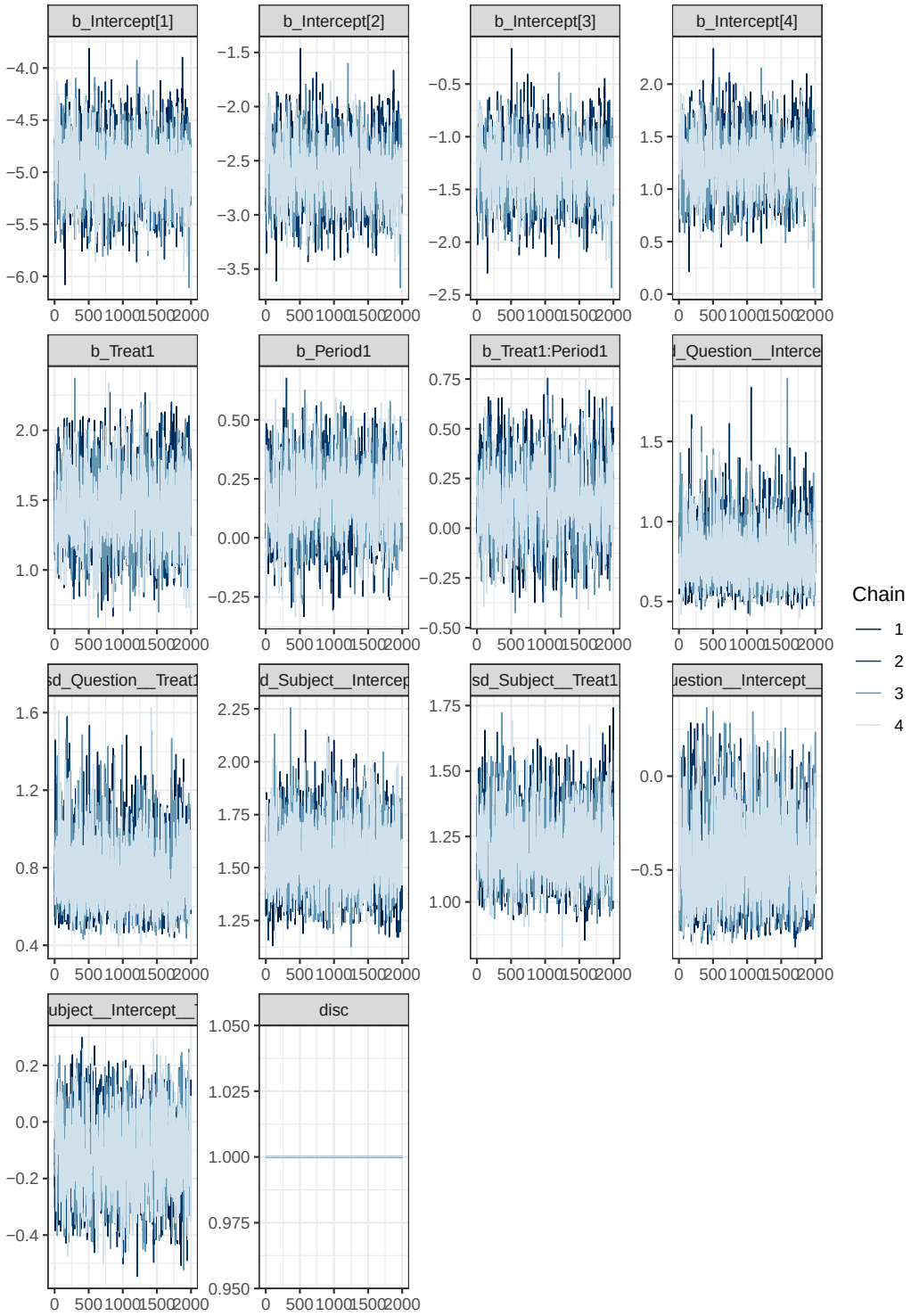



Figura 4.13: Cadenas MCMC del modelo Response ~ Treat * Period + (1 + Treat | Subject) + (1 + Treat | Question).

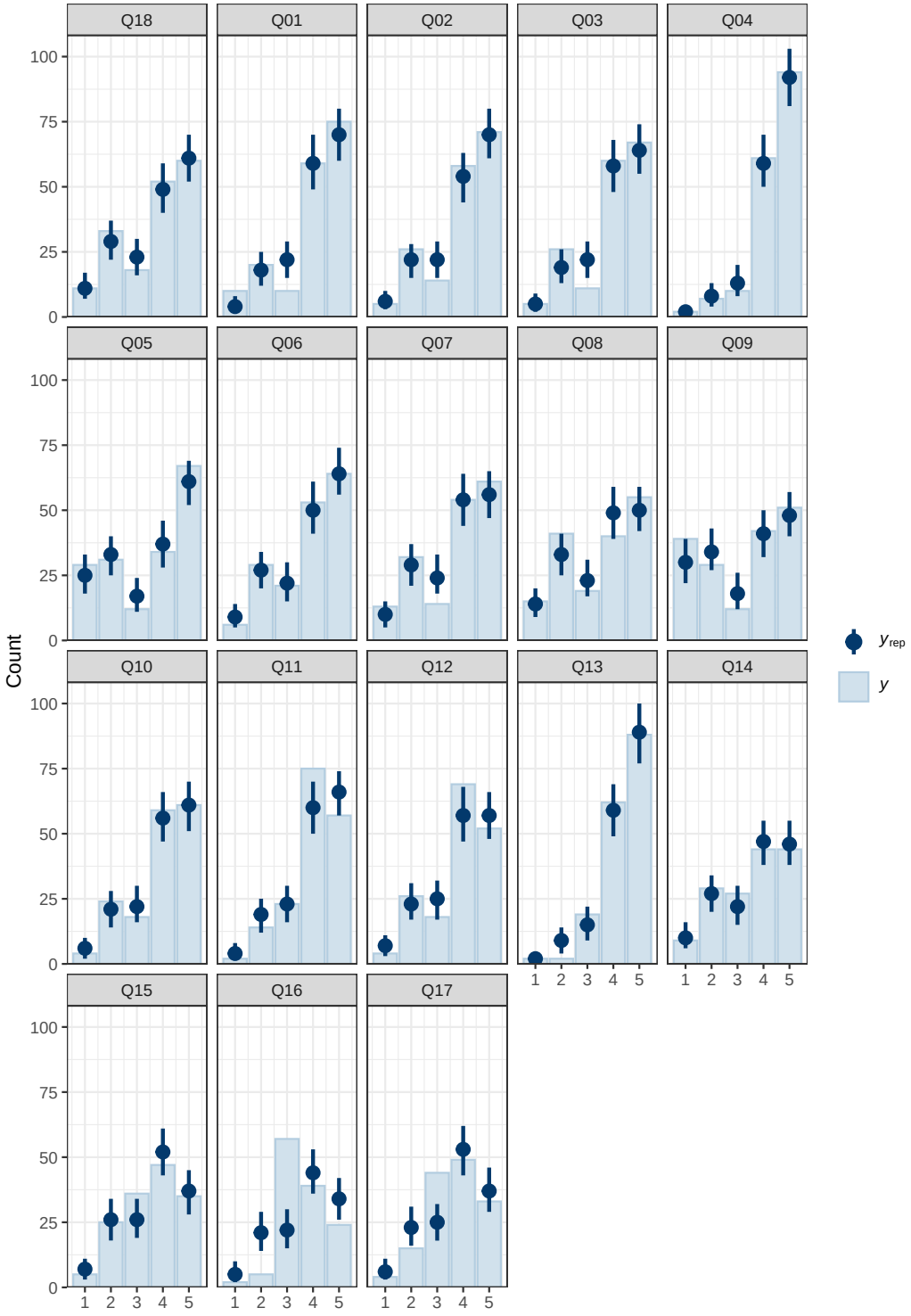



Figura 4.14: Comparación de los valores reales con los obtenidos a partir de la función predictiva a posteriori del modelo Response ~ Treat * Period + (1 + Treat | Subject) + (1 + Treat | Question).

Capítulo 5

# Resultados

En esté capítulo se comentarán los resultados de las técnicas estadísticas y de los modelos propuestos (ver Capítulo 4).

## Correlación entre preguntas con el alfa de Cronbach

El coeficiente `alfa de Cronbach` (ver Sección 2.2) de la escala de Likert es 0.921 lo que indica una muy buena correlación entre las respuestas a todas los ítems. Este valor apenas se ve alterado si se elimina una de ellos. Esto nos permite concluir que todos los ítems pertenecen a la misma escala de Likert. En la Tabla 5.1 mostramos los ítems que más contribuyen al índice `alpha de Cronbach`. Un resultado interesante es constatar que el ítem $Q18$, que es la valoración general del cuestionario, es el que mejor contribución tiene al índice.

Cuadro 5.1: Relación de cada ítem con el índice alpha de Cronbach.

(a) Variables con mayor asociación.

| Q18 | Q05 | Q06 | Q09 | Q08 | Q07 | Q10 | Q12 | Q02 |
|---|---|---|---|---|---|---|---|---|
| 0.86 | 0.81 | 0.79 | 0.79 | 0.77 | 0.75 | 0.73 | 0.72 | 0.71 |

(a) Variables con menor asociación.

| Q03 | Q14 | Q01 | Q11 | Q15 | Q13 | Q04 | Q16 | Q17 |
|---|---|---|---|---|---|---|---|---|
| 0.68 | 0.66 | 0.65 | 0.64 | 0.56 | 0.51 | 0.46 | 0.44 | 0.42 |

Cuadro 5.4: Tablas de contingencia.

(a) ~ Response + Treat

| Response | A | B |
|---|---|---|
| 1 | 4 | 163 |
| 2 | 55 | 359 |
| 3 | 135 | 248 |
| 4 | 548 | 409 |
| 5 | 755 | 304 |

(b) ~ Response + Period

| 1 | 2 |
|---|---|
| 78 | 89 |
| 199 | 215 |
| 198 | 185 |
| 440 | 517 |
| 572 | 487 |

(c) ~ Response + Seq

| AB | BA |
|---|---|
| 89 | 78 |
| 210 | 204 |
| 192 | 191 |
| 375 | 582 |
| 600 | 459 |

## Asociación de variables con la prueba de homogeneidad $X^2$

En la Sección 2.3 se describe el marco teórico de esta prueba no paramétrica. Se ha contrastado la existencia de asociación entre la variable respuesta *Response* y cada una de las tres variables más importantes de nuestro modelo (*Treat*, *Period* y *Seq*). En la Tabla 5.4 se muestran las tablas de contingencia. La Tabla 5.5 permite comprobar que todos los contrastes son significativos, con lo que se rechaza la hipótesis nula de homogeneidad. Todas las variables explicativas tienen influencia en la variable respuesta.

Cuadro 5.5: Valores del contraste de hipótesis $\chi^2$

| | ~ Response + Treat | ~ Response + Period | ~ Response + Seq |
|---|---|---|---|
| $\chi^2$ | 621.497 | 15.039 | 64.904 |
| $p$-value | $4.5778215 \times 10^{-132}$ | 0.0101975 | $1.1733569 \times 10^{-12}$ |
| $df$ | 5 | 5 | 5 |

## Comparación con *Odds Ratio*

En Sección 2.3 y en Sección 4.2 se exponen el marco teórico y la fundamentación de los contrastes de hipótesis realizados respectivamente. El contraste de hipótesis del *log OR* del subtitulado para cada grupo no produce significación estadística en ningún nivel de respuesta, por lo que según esta prueba estadística la secuencia de subtitulado no influiría en la respuesta de los estudiantes (ver Tabla 5.6). Es decir, de acuerdo a este test la secuencia no influye en la valoración de los subtítulos.

Cuadro 5.6: $LogOR \sim$ Treat + Seq + Response_l

| 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 realizamos este contraste entre subtítulos y periodos podemos constatar la existencia de un efecto periodo de signo contrario para las preguntas 4 y 5 (ver Tabla 5.7).

Cuadro 5.7: Log OR ~ Treat + Period + Response_l

| 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 |

## Modelado

El modelo de efectos mixtos ajustado con Regresión Logística:

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

estima la probabilidad de que la respuesta a un ítem en el subtitulado $A$ sea mejor que en el subtitulado $B$ frente a que sea igual o peor. El intercepto del modelo ajustado es 0.404. Por ello, la probabilidad de que se otorgue una mayor puntuación en $A$ que en $B$ es del 59.95%. En la Tabla 5.8 se muestra la probabilidad de que la respuesta sea mejor en $A$ que en $B$ por ítem.

Cuadro 5.8: Probabilidad de que la respuesta a un ítem sea A > B frente a A <= B

| Question | Prob |
|---|---|
| Q18 | 0.70 |
| Q01 | 0.36 |
| Q02 | 0.57 |
| Q03 | 0.38 |
| Q04 | 0.16 |
| Q05 | 0.81 |
| Q06 | 0.68 |
| Q07 | 0.58 |
| Q08 | 0.68 |
| Q09 | 0.86 |
| Q10 | 0.45 |
| Q11 | 0.40 |
| Q12 | 0.50 |
| Q13 | 0.24 |
| Q14 | 0.64 |

| | |
|---|---|
| Q15 | 0.30 |
| Q16 | 0.24 |
| Q17 | 0.33 |

En la Sección 4.2 se evaluaron distintas parametrizaciones de la Regresión Ordinal Acumulativa tanto desde el punto de vista frecuentista como bayesiano y 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 es el de la Ecuación 4.2, que se reproduce aquí en sintaxis R:

> Response ~ Treat*Period + (1 + Treat | Subject) + (1 + Treat | Question)

Este modelo produce en los dos paradigmas unas estimaciones similares, como se puede comprobar en la Tabla 5.9.

Cuadro 5.9: Comparación frecuentista/bayesiano de coeficientes estimados en el modelo Response ~ Treat * Period + (1 + Treat | Subject) + (1 + Treat | Question).

| | ordinal::clmm | | | brms::brm | | |
|---|---|---|---|---|---|---|
| Name | Estimation.clmm | conf.2.5% | conf.97.5% | Estimation.brm | cred.2.5% | cred.97.5% |
| 1\|2 | -5.00 | -5.52 | -4.49 | -4.96 | -5.50 | -4.42 |
| 2\|3 | -2.65 | -3.12 | -2.18 | -2.60 | -3.09 | -2.11 |
| 3\|4 | -1.37 | -1.83 | -0.90 | -1.32 | -1.80 | -0.82 |
| 4\|5 | 1.18 | 0.72 | 1.65 | 1.24 | 0.76 | 1.74 |
| Treat1 | 1.43 | 1.02 | 1.84 | 1.45 | 1.01 | 1.93 |
| Period1 | 0.17 | -0.09 | 0.43 | 0.17 | -0.10 | 0.43 |
| Treat1:Period1 | 0.14 | -0.18 | 0.46 | 0.14 | -0.19 | 0.47 |
| Question.sd(Intercept) | 0.70 | | | 0.76 | 0.54 | 1.13 |
| Question.sd(Treat1) | 0.68 | | | 0.74 | 0.53 | 1.12 |
| Subject.sd(Intercept) | 1.49 | | | 1.53 | 1.30 | 1.83 |
| Subject.sd(Treat1) | 1.17 | | | 1.21 | 1.02 | 1.45 |
| Question.cor(Intercept,Treat1) | -0.53 | | | -0.48 | -0.79 | 0.00 |
| Subject.cor(Intercept,Treat1) | -0.13 | | | -0.11 | -0.36 | 0.13 |

Capítulo 6

# Discusión y conclusiones

## 6.1 Comparación con *Odds Ratio*

En la Sección 5 se constató la existencia de un efecto periodo en las respuestas 4 y 5. El test es significativo porque el ratio entre subtítulos de respuestas con valor 4 es diferente en cada periodo habiendo mayor 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 6.1 permite una comprobación visual. Esto estaría indicando que los estudiantes de ambos grupos prestaron más atención o fueron más exigentes en el segundo visionado y decidieron no otorgar la puntuación máxima. 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 provocan que el resto de niveles pierdan o ganen respectivamente la misma cantidad. En cualquier caso, vemos que 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 que cualitativamente es poco importante ya que las respuestas 4 y 5 son ambas valoraciones positivas.

## 6.2 Modelado

La probabilidad de que los alumnos puntúen mejor el subtitulado $A$ que el $B$ frente a que lo puntúen igual o peor se estima que es del 59.95%. Sin embargo, la distribución por ítems no es uniforme (ver Tabla 5.8). Así, hay ítems en que esta probabilidad sobrepasa el 80% (Q09, Q05). Sin embargo

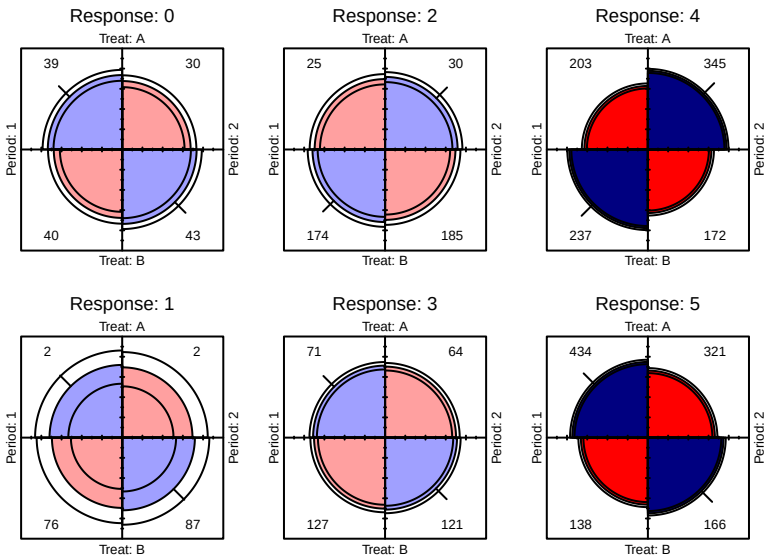



Figura 6.1: OR ~ Treat + Period + Response

hay otros en que no supera el 25% (Q04, Q16, Q13). La respuesta al ítem Q18 está por encima de la media (69.82%). Parece claro que los ítems en los que la respuesta de los estudiantes es diferente en los dos subtitulados es porque aprecian diferencias reales entre ellos. Quedaría por dilucidar si en los ítems en los que no hay diferencia en las respuestas, se debe a que realmente no las hay o si es que los estudiantes no son capaces de detectarlas.

La magnitud del efecto del subtitulado sobre las valoraciones es muy superior a la de los efectos periodo y secuencia (ver Tabla 5.9), por lo que podemos realizar interpretaciones del modelo sin que nos deba preocupar su existencia. En la Figura 6.2 se representan 50 muestras de la esperanza de la distribución predictiva a posteriori para cada pregunta y nivel de subtitulado marginalizadas por periodo y estudiante. La primera conclusión que podemos extraer es que el modelo tiene bastante incertidumbre sobre los valores de respuesta a cada pregunta no superando casi nunca el 50% de probabilidad para todas las preguntas y niveles de subtitulado. En general se observa en la mayoría de las preguntas del nivel de subtitulado $A$ que los alumnos están bastante seguros de que la respuesta a las preguntas 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 pregunta podemos extraer las siguientes conclusiones:

- En las preguntas $Q04$ y $Q13$ 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 las preguntas $Q15$, $Q16$ y $Q17$, la opción de respuesta más probable es 4 en ambos subtitulados. 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 preguntas $Q01$, $Q02$, $Q03$, $Q10$, $Q11$ y $Q12$ 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.

- Las preguntas $Q06$, $Q07$, $Q14$ y $Q18$ no son muy diferentes de las 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.

- Las preguntas $Q05$, $Q08$ y $Q09$ son las que más diferencias entre subtitulados presentan. La respuesta más probable en el subtitulado $A$ es 5 (en $Q08$ y en $Q09$ 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 $Q05$ y en $Q09$ la segunda respuesta más probable al subtitulado $B$ es 1 y 4 en la $Q08$.

En definitiva, nuestro modelo predice que los estudiantes están bastante de acuerdo en que en las preguntas $Q05$ y $Q09$ hay una diferencia de calidad importante entre subtitulados. También están de acuerdo en que en las preguntas $Q04$ y $Q13$ no hay apenas cambio entre los subtitulados. En las preguntas $Q15$, $Q16$ 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.

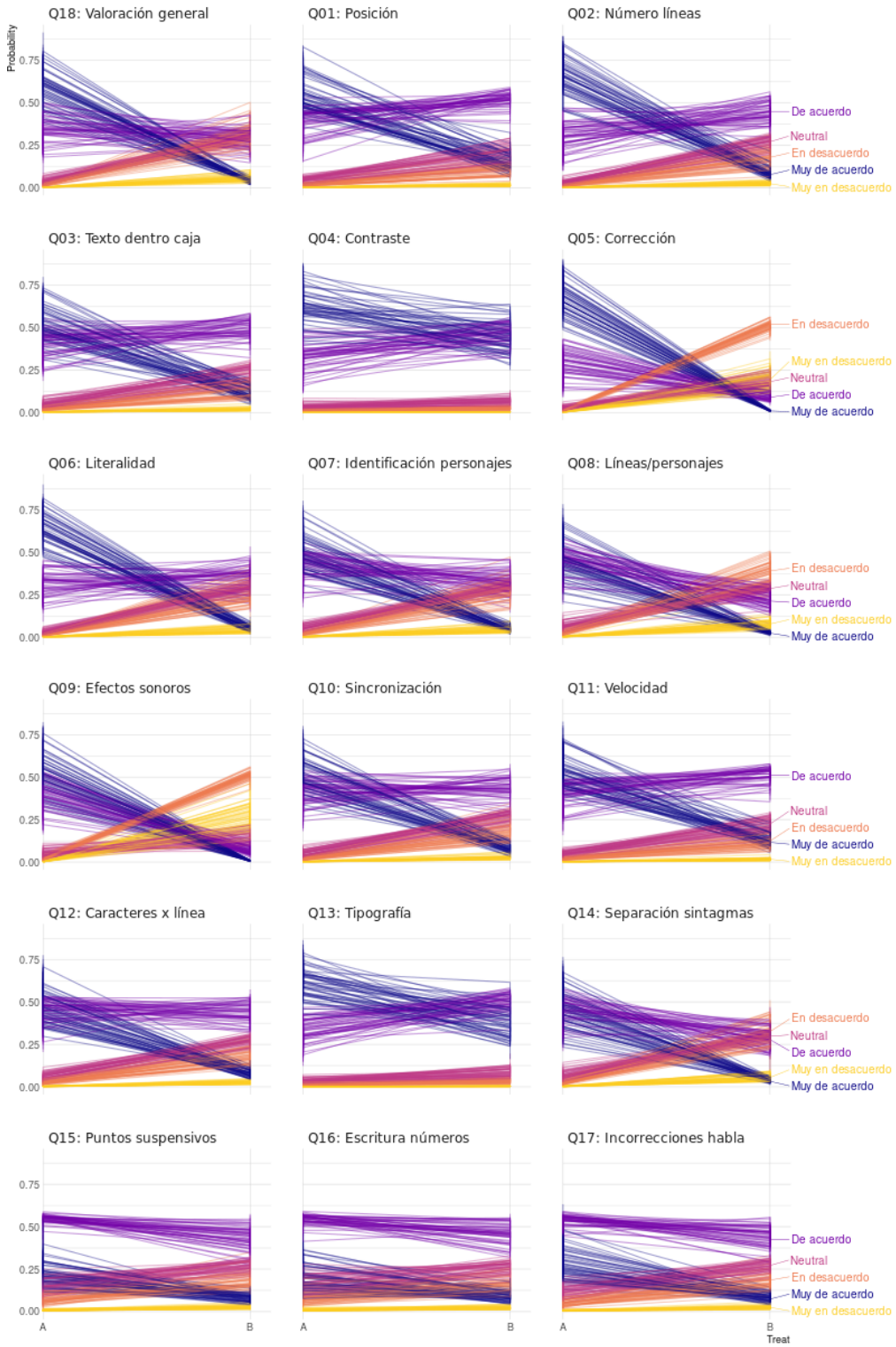



Figura 6.2: Muestreo de la función predictiva a posteriori por tratamiento y pregunta del modelo Response ~ Treat * Period + (1 + Treat | Subject) + (1 + Treat | Question).

# Referencias


AENOR (2012). *UNE 153010 Subtitulado para personas sordas y personas con discapacidad auditiva*. Asociación Española de Normalización y Certificación (vid. pág. 18).

Agresti, A. (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. 6, 7).

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. 14).

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. 36).

Bürkner, P.-C. (nov. de 2021). «Bayesian Item Response Modeling in R with brms and Stan». En: *Journal of Statistical Software* 100. DOI: 10.18637/jss.v100.i05 (vid. pág. 42).

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. 9, 42, 49).

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. 12, 13).

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. pág. 35).

Christensen, R. H. B. (2018). «Cumulative Link Models for Ordinal Regression with the R Package ordinal». En: (vid. pág. 36).

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. págs. 6, 8).

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari y D. Rubin (nov. de 2013). *Bayesian Data Analysis*. DOI: 10.1201/b16018 (vid. pág. 12).

Harrell, F. (2020). «Violation of Proportional Odds is Not Fatal». En: URL: https://www.fharrell.com/post/po/ (vid. pág. 11).

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. pág. 35).

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. 3).

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. 4).

Liu, X. (abr. de 2022). *Categorical Data Analysis and Multilevel Modeling Using R*. Ed. por S. P. Ltd. (vid. págs. 11, 12).

Lui, K.-J. (ago. de 2016). *Crossover Designs: Testing, Estimation, and Sample Size*. DOI: 10.1002/9781119114710 (vid. pág. 4).

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. pág. 18).

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. 13).

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. 18).

Schweinberger, M. (2020). *Questionnaires and Surveys: Analyses with R*. 2020/12/11. https://slcladal.github.io/survey.html. The University of Queensland, Australia. School of Languages y Cultures. Brisbane (vid. pág. 5).

Senn, S. (2022). *Cross-over Trials In Clinical Research*. Ed. por L. John Wiley. DOI: 10.1002/0470854596 (vid. pág. 4).

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. 35).

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. pág. 35).

Apéndice A

# Efecto secuencia e interacción tratamiento vs. periodo.

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

## A.1 Preparación.

Partimos 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
2   0.13153117     A      1  AB
3  -0.07891709     A      2  BA
```

[1] Obsérvese que se 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.

```
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
```

Calculamos las medias por cada nivel de factor y combinaciones de niveles que utilizaremos luego 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()
```

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

| X.Intercept. | TreatB |
|---|---|
| 0.03624217 | -0.03966751 |

Vemos que el intercepto es la media de la respuesta en el nivel de tratamiento $A$:

```
mTreat[1]
```

```
        A
0.03624217
```

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
```

Podemos ver la matriz ampliada añadiendo el intercepto, que siempre será 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, multiplicamos el valor del contraste por el parámetro. Así:

- De la primera fila obtenemos que el efecto del tratamiento $A$ es el intercepto: $A = 1 \cdot Intercept + 0 \cdot TreatB$.
- De la segunda fila obtenemos 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()
```

Cuadro 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 |

Vemos 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:

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

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

```
  (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 concluimos 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 hemos calculado de forma diferente $TreatB : Period2$. Podemos aplicar la fórmula anterior y comprobar 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.

Vamos a incorporar 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 vemos 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()
```

Cuadro 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 estamos usando el contraste con `codificación de tratamientos`. Pero si cambiamos a `codificación de sumas`:

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

Y volvemos a ajustar los modelos que ya usarán el `contraste suma`, podemos comprobar 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()
```

Cuadro 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()
```

Cuadro 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 contrastes es diferente. Para explicarlo, mostramos 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"
```

Vemos que 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 interceptor es la media de la media de cada uno de los niveles de factor. ¿Por qué?. El interceptor 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. Podemos comprobar que la media global coincide con el interceptor 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
```

```
          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.

[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.

```
(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
```

- Podemos 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
```