Universidad de Oviedo

ESCUELA POLITÉCNICA DE INGENIERÍA DE GIJÓN

GRADO EN CIENCIA E INGENIERÍA DE DATOS

ÁREA DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

ANÁLISIS DE DATOS LONGITUDINALES

D. Pablo Álvarez Arnedo
TUTORES: D. Carlos De la Calle Arroyo
Dña Arís Fanjul Hevia

JULIO DE 2025

Tabla de contenidos

1 Introducción

Estas últimas décadas, muchos gobiernos y organizaciones internacionales han integrado indicadores de bienestar en sus sistemas y procedimientos estadísticos, considerando que las métricas tradicionales como el Producto Interior Bruto (PIB) no son suficientes para determinar el progreso de su sociedad. En este contexto, surge el World Happiness Report (Helliwell et al. 2024), una iniciativa impulsada por las Naciones Unidas que, desde 2012, proporciona una evaluación anual del nivel de felicidad de los países a partir de encuestas a sus ciudadanos y distintos objetivos relacionados con factores económicos, sociales y políticos.

Este interés por determinar y comprender la felicidad plantea distintos retos analíticos. A diferencia de otros indicadores, la felicidad presenta una fuerte variabilidad individual y temporal, lo que exige metodologías estadísticas capaces de capturar tanto las diferencias entre países como las evoluciones temporales dentro de cada uno. De esta manera, la construcción de modelos longitudinales y estructuras jerárquicas se postula como un planteamiento adecuado para estudiar este hecho de forma rigurosa.

Los datos longitudinales han desempeñado un papel fundamental en el análisis de sucesos que evolucionan a lo largo del tiempo. Aunque se llevan usando desde el siglo XIX en estudios de crecimiento infantil o en registros médicos hospitalarios, fue durante el siglo XX cuando se consolidaron como una herramienta estadística esencial en disciplinas como la epidemiología, la psicología o las ciencias sociales. A diferencia de los datos transversales, los datos longitudinales permiten observar cómo cambian las observaciones (individuos, países, instituciones) a lo largo del tiempo, lo que permite sacar conclusiones más precisas sobre causalidad, evolución y efectos individuales. Estudiar datos longitudinales también implica afrontar ciertos desafíos, como la dependencia y correlación entre medidas repetidas, la gestión de datos faltantes o la necesidad de modelos que integren múltiples niveles de variación. Este trabajo aplica este planteamiento a través de técnicas modernas de modelado a una base de datos compleja y rica en dimensiones temporales y jerárquicas.

1.1 Objetivo del trabajo

El objetivo principal de este Trabajo de Fin de Grado es estudiar y aplicar técnicas estadísticas avanzadas adecuadas para el análisis de datos longitudinales, enfocándonos en los modelos mixtos tanto lineales como generalizados. A través de un caso práctico concreto como lo es la evolución del índice global de felicidad a nivel mundial, se busca mostrar cómo estas

herramientas son capaces de modelar estructuras jerárquicas, capturar tendencias temporales y realizar inferencias sólidas en contextos donde las observaciones están organizadas en múltiples niveles. Aunque el análisis se centra en los datos del World Happiness Report, enriquecidos con variables políticas, este método se puede aplicar en diferentes situaciones en las que los datos presentan una estructura longitudinal o multinivel. Además, se desarrolla una herramienta interactiva que facilita la visualización, modelización y predicción de este tipo de sucesos, fomentando así la reproducibilidad y accesibilidad del análisis.

Específicamente, este trabajo plantea integrar diversas fuentes de datos sobre felicidad, condiciones socioeconómicas y políticas; aplicar técnicas de análisis exploratorio para identificar posibles patrones en la evolución de la felicidad; ajustar modelos mixtos que respeten la estructura jerárquica de los datos (países, regiones, años); y evaluar el ajuste y la validez de dichos modelos utilizando criterios estadísticos adecuados. Una aportación clave del proyecto es el desarrollo de una aplicación interactiva con Shiny (Chang et al. 2024), una librería del lenguaje de programación R (R Core Team 2024a), que permite contrastar los análisis presentados con anterioridad, explorar distintas configuraciones de modelos y generar predicciones del Happiness Score de forma accesible, reproducible y visualmente intuitiva. Esta herramienta no solo facilita la comprensión de los resultados, sino que también generaliza el uso de técnicas estadísticas avanzadas para un público más amplio.

El trabajo adopta una perspectiva cuantitativa, basada en el uso de técnicas estadísticas para modelar datos longitudinales. Se parte de una base de datos principal (World Happiness Report 2015–2024), a la que se añaden variables políticas obtenidas de fuentes como Freedom in the World y Democracy Data. A nivel metodológico, el análisis se estructura en tres niveles: una exploración inicial de los datos, que incluye limpieza, imputación de valores faltantes, análisis de outliers y visualizaciones; el ajuste de modelos estadísticos, utilizando modelos de regresión, modelos lineales mixtos (LMM) y modelos lineales generalizados mixtos (GLMM), con efectos aleatorios por país y, en algunos casos, por región; y, finalmente, la construcción de una herramienta interactiva, mediante Shiny, que engloba todo el proceso de análisis, desde la exploración hasta la validación y predicción de los modelos.

Este enfoque permite capturar tanto las diferencias estructurales entre países como las tendencias temporales de cada uno de ellos, facilitando la interpretación de los resultados.

1.2 Estructura del trabajo

El contenido del trabajo se organiza en siete capítulos, que se resumen a continuación:

- Capítulo 2 – Datos longitudinales: se introduce el concepto de datos longitudinales, sus características específicas, y se justifica la necesidad de utilizar modelos mixtos en lugar de técnicas de estadística clásicas.

- Capítulo 3 – Modelos mixtos: se presenta el marco teórico de los modelos lineales mixtos (LMM) y modelos lineales generalizados mixtos (GLMM), incluyendo su formulación, métodos de estimación, validación y predicción.
- Capítulo 4 – Análisis exploratorio de la base de datos: se describen las tareas de limpieza, integración y análisis inicial de los datos del World Happiness Report, complementados con variables políticas de otras bases de datos.
- Capítulo 5 – Construcción de modelos predictivos: se proponen distintos métodos para la construcción de modelos, como una estrategia combinada top-down y bottom-up para ajustar modelos clásicos que expliquen el Happiness Score, mientras se evalúan distintas combinaciones de variables y efectos. Se identifican modelos válidos, se analizan sus coeficientes, y se realizan predicciones para el año 2025.
- Capítulo 6 – Aplicación Shiny para la modelización de la felicidad: se describe detalladamente la aplicación interactiva desarrollada con Shiny, que permite realizar el análisis completo (exploración, modelización, validación, predicción) desde una interfaz accesible.
- Capítulo 7 – Conclusiones y mejoras futuras: se realiza una valoración crítica del trabajo, identificando las principales aportaciones metodológicas y prácticas, y proponiendo posibles líneas de mejora y extensión para futuros trabajos.

Desde una perspectiva académica, el trabajo consiste en una aplicación práctica y completa del análisis de datos longitudinales, integrando técnicas de modelización, visualización, validación y desarrollo. Permite afianzar conocimientos adquiridos durante el Grado en Ciencia e Ingeniería de Datos, en especial en estadística, investigación y desarrollo en R.

Desde el punto de vista social, el trabajo aborda una cuestión de alto interés como es la felicidad global y la calidad de vida. Al ofrecer una herramienta interactiva para explorar los factores influyentes de la felicidad en distintos ambientes temporales y geográficos, el proyecto puede resultar útil para investigadores, docentes, periodistas o responsables políticos interesados en promover el bienestar en sus comunidades.

2 Datos longitudinales

2.1 Datos con medidas repetidas

Los **datos longitudinales** son aquellos que obtenemos al realizar distintas medidas a un mismo individuo (personas, regiones, etc.). Dichas medidas se pueden observar repetidamente a lo largo del tiempo (análisis temporal), como el salario anual de diferentes personas a lo largo de varios años; del espacio (análisis espacial), por ejemplo, al medir la contaminación del aire de distintas ciudades en un mismo día; o a lo largo del espacio y tiempo (análisis espacio-temporal), como puede ser la monitorización de la expansión de una enfermedad en distintas regiones a lo largo del tiempo. Como lo más frecuente es encontrar medidas repetidas en el tiempo, consideraremos ese caso sin perder generalidad alguna, ya que todo lo presentado se puede aplicar a los otros dos casos. Por esto, a los datos longitudinales también se les conoce como medidas repetidas.

Tal y como se expone en *Curso de datos longitudinales* (Subirana 2020), los datos longitudinales combinan características de las series temporales y los estudios transversales, lo que exige técnicas específicas de análisis que tengan en cuenta la dependencia entre observaciones repetidas de la misma unidad. De forma similar, tanto *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* (Faraway 2006) como *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R* (Roback y Legler 2021) destacan la necesidad de modelos con estructuras jerárquicas que puedan captar tanto la variación entre unidades como dentro de ellas.

El análisis de este tipo de medidas nos permite detectar cambios o tendencias temporales en nuestras variables, lo cual nos puede llevar a observar patrones díficiles de contemplar usando otro tipo de técnicas. Es común usar este tipo de datos en estudios donde se busca evaluar cómo evolucionan ciertas características o mediciones bajo distintas condiciones. En el ámbito biosanitario, los datos longitudinales son fundamentales para investigar la progresión de enfermedades, la eficacia de tratamientos y el impacto de intervenciones médicas. En este capítulo, exploraremos las características clave de los datos longitudinales y profundizaremos en las razones por las que los métodos clásicos, como la regresión lineal simple, no deben aplicarse a este tipo de datos.

Como ya hemos mencionado anteriormente, una de las características que definen a los datos longitudinales es que tenemos medidas repetidas del mismo sujeto a través de diferentes observaciones. No obstante, dichas observaciones no están organizadas de cualquier manera, sino que están agrupadas por unidades (pacientes o regiones). Todo ello significa que cada

unidad tiene varias observaciones en diferentes momentos temporales, haciendo que los datos longitudinales adopten una estructura jerárquica.

Esta estructura nos lleva a asumir una de las claves en todo este proceso, la dependencia entre las observaciones, la cual nos indica que las medidas dentro de la misma unidad tienden a estar correlacionadas. También tenemos que destacar las distintas variables que definen a dichos datos, que suelen clasificarse según diferentes características. Como la mayoría de mediciones se realizan en distintos instantes de tiempo, es normal que su valor cambie a lo largo del tiempo, permitiendo considerarlas como variables tiempo-dependientes, lo que significa que sus cambios pueden estar relacionados con el tiempo y pueden ser modeladas para entender tendencias o patrones; pero también hay que tener en cuenta que hay otras variables que cambian igual en el tiempo para todos los sujetos (como el régimen político del país) que no consideraremos tiempo-dependientes y otras que directamente son constantes como el sexo.

El análisis de datos longitudinales se centra en aprovechar las medidas repetidas para tratar cuestiones que no pueden ser respondidas adecuadamente con otros tipos de datos. Uno de los principales objetivos del análisis de estos datos es observar la evolución de una variable a lo largo del tiempo, lo cual nos permitiría poder detectar si los cambios de las variables siguen ciertos patrones que tendríamos que tener en cuenta en el análisis. Esta identificación de patrones nos puede aportar información y conocimientos clave, ya que nos ayuda a formular ciertas hipótesis que nos orientan hacia una visión concreta. Otra parte importante reside en comparar si la evolución de una variable a lo largo del tiempo es igual para distintas partes de la población, y ver si existen factores que determinan la evolución de dicha variable, en cuyo caso deberíamos estudiar cómo dichos factores interactúan en el tiempo.

Los datos longitudinales tienen aplicaciones en una gran diversidad de áreas, ya que el estudio de medidas a lo largo del tiempo está presente en diferentes ámbitos. Por ejemplo, los datos longitudinales tienen una gran importancia en el ámbito biosanitario, como puede ser en pruebas donde hay medidas repetidas de presión arterial en un grupo de pacientes durante un tratamiento donde se puede monitorear la salud de los pacientes para evaluar la efectividad del tratamiento. Además, este tipo de datos también tiene su relevancia en otras áreas como la educación; por ejemplo, la evaluación de las puntuaciones de un estudiante a lo largo de varios exámenes anuales puede destacar posibles áreas de mejora por parte del alumno o algunas estrategias didácticas implementables por parte del profesorado. Otra de las áreas en la que los datos longitudinales juegan un papel clave es en la alimentación, mediante el estudio de diferentes dietas a diferentes grupos de la población a lo largo del tiempo a través de medidas como la actividad física, peso corporal, etc. y cómo estas rutinas aportan ciertos beneficios o riesgos a la salud de los individuos. En otros ámbitos como en el marketing también encontramos casos en los que se utilizan datos longitudinales, como son encuestas de opinión realizadas periódicamente a las mismas personas que pueden ser de gran ayuda a la hora de evaluar posibles campañas de concienciación, o simplemente estudiar el comportamiento y la opinión de la población. Además, los datos longitudinales juegan un papel clave en el estudio de aspectos sociales, políticos y demográficos. Un ejemplo es el análisis de la felicidad y bienestar de los países a lo largo del tiempo, lo que permite identificar cómo factores como

el crecimiento económico, la percepción de la corrupción, la generosidad y el apoyo social influyen en la felicidad de la población. Estos estudios pueden ser fundamentales para que los gobiernos establezcan políticas que promuevan un mayor nivel de calidad de vida y bienestar social. También, en el ámbito demográfico, los datos longitudinales pueden ayudar a analizar la evolución de indicadores clave como la esperanza de vida o la migración en diferentes regiones del mundo, proporcionando información determinante a la hora de tomar de decisiones a nivel global.

A pesar de su gran utilidad, los datos longitudinales presentan varias complicaciones. En primer lugar, aunque las mediciones suelen realizarse en intervalos de tiempo predeterminados, no siempre disponemos de todas las observaciones esperadas debido a la presencia de valores faltantes. Estos valores faltantes pueden ser producto de la ausencia de un paciente en una consulta médica, la falta de respuesta en una encuesta periódica o errores en la recolección de datos. Además, en muchos estudios, los individuos no siempre son medidos en los mismos instantes de tiempo, por lo que podemos no tener el mismo número de medidas repetidas por individuo, lo que lleva a una estructura desigual en los datos que debe ser tratada con técnicas adecuadas. Estas dificultades pueden generar desafíos en el modelado y en la comparación de diferentes evoluciones, por lo que es fundamental aplicar estrategias estadísticas como imputación de valores faltantes, modelado con efectos aleatorios o técnicas para datos desbalanceados. Según Isaac Subirana en su *Curso de datos longitudinales* (Subirana 2020), los modelos lineales mixtos proporcionan una herramienta útil para abordar estos problemas, permitiendo modelar la estructura de correlación y manejar la variabilidad de las observaciones. Esto se puede apreciar en la Figura 2.1, donde tenemos por un lado intervalos regulares, irregulares y con datos faltantes:

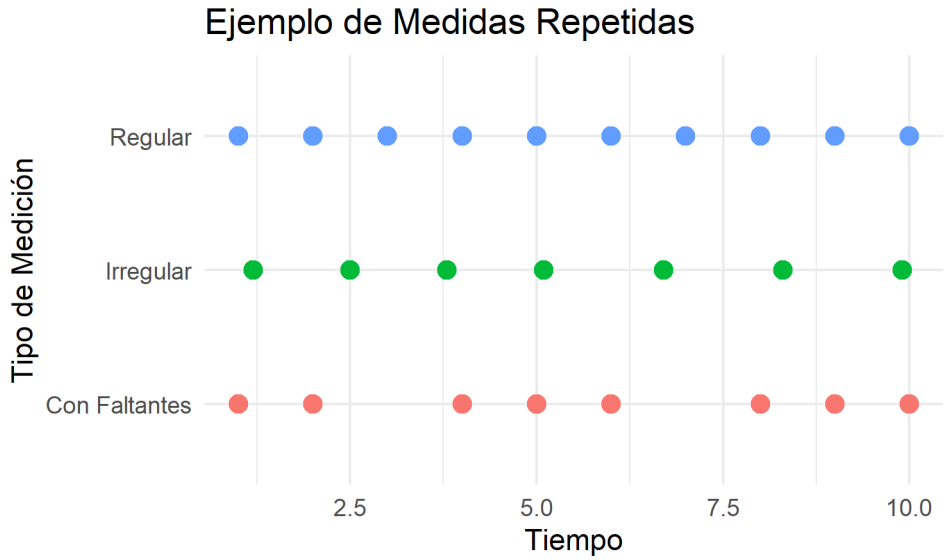

Figura 2.1: Ejemplo de medidas repetidas en diferentes estructuras temporales.

Como podemos apreciar en la Figura 2.1, tenemos por un lado intervalos regulares en los que las mediciones se toman a intervalos de tiempo predefinidos, intervalos regulares con valores ausentes en los que se han perdido algunas medidas a lo largo del tiempo, y, por último, intervalos irregulares en los que las mediciones no siguen una periodicidad fija. Estas complicaciones pueden suponer un problema, y es importante tenerlas en cuenta.

2.2 Conceptos básicos de la regresión lineal simple

La **regresión lineal simple** es un método estadístico utilizado para modelar la relación entre una variable dependiente Y (respuesta) y una variable independiente X (predictora) mediante una ecuación lineal. El modelo se define matemáticamente de la siguiente manera:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

donde:

- Y representa la variable dependiente (respuesta).
- X es la variable independiente (predictora).
- β_0 es el intercepto, que indica el valor esperado de Y cuando $X = 0$.
- β_1 es la pendiente, que mide el cambio esperado en Y por cada unidad de cambio en X.
- ε representa el término de error, que captura la variabilidad no explicada por el modelo.

Para que la regresión lineal simple sea válida y produzca estimaciones fiables, deben cumplirse ciertos supuestos fundamentales:

1. **Linealidad:** la relación entre la variable independiente X y la dependiente Y debe ser lineal, lo que significa que un cambio en X se traduce en un cambio proporcional en Y.
2. **Independencia:** las observaciones deben ser independientes entre sí; es decir, los valores de Y no deben estar correlacionados con otras observaciones.
3. **Normalidad de los errores:** se asume que los errores ϵ_i siguen una distribución normal con media cero ($\varepsilon_i \sim N(0, \sigma^2)$), lo cual es importante para sacar conclusiones sobre los coeficientes β_0 y β_1.
4. **Homocedasticidad:** la varianza de los errores debe ser constante para todos los valores de X; es decir, la dispersión de los valores de Y en torno a la línea de regresión debe ser uniforme.

Cuando se satisfacen los supuestos del modelo, la regresión lineal simple permite obtener estimaciones sólidas y no sesgadas de los parámetros que describen la relación entre las variables. Además, esta técnica permite realizar contrastes de hipótesis o construcción de intervalos de confianza, para evaluar la significancia estadística del efecto de la variable independiente sobre la dependiente.

2.3 ¿Por qué no se puede usar la estadística clásica?

La estadística clásica, como la regresión lineal simple, parte de la suposición fundamental de que todas las observaciones son independientes entre sí. Sin embargo, en datos longitudinales, esta independencia no se cumple debido a la correlación entre medidas repetidas de la misma unidad a lo largo del tiempo. Los datos longitudinales presentan ciertas características que precisan de métodos estadísticos más avanzados.

Uno de los principales desafíos, ya mencionado anteriormente, es la dependencia entre observaciones, ya que los datos recogidos de un mismo individuo suelen estar correlacionados, lo que genera un patrón estructurado que no es capturado por modelos clásicos. Esta correlación también afecta a la estructura de los errores, ya que las medidas repetidas pueden estar influenciadas por factores externos o por variables no observadas, lo que genera una relación entre los errores que los modelos clásicos no pueden modelar correctamente. Además, la variabilidad entre individuos es un aspecto clave en datos longitudinales, ya que no todos los sujetos presentan la misma evolución en el tiempo. Los modelos clásicos suelen asumir una varianza constante, lo cual no es adecuado en este contexto ya que no permite capturar diferencias individuales ni estructuras de correlación complejas.

Todos estos factores hacen que el uso de modelos estadísticos clásicos, como la regresión lineal simple, no sea adecuado para el análisis de datos longitudinales. En su lugar, es necesario recurrir a ciertos métodos, como los modelos lineales mixtos, que permiten modelar tanto los efectos fijos como los efectos aleatorios para capturar adecuadamente la variabilidad y dependencia propia de estos datos. La mejor manera de comprender estas limitaciones es a través de un ejemplo práctico.

2.3.1 Ejemplo conceptual

Para ilustrar las limitaciones de la estadística clásica en el análisis de datos longitudinales, vamos a considerar un conjunto de datos sobre ingresos anuales (en euros) de 10 personas medidos a lo largo de varios años. Vamos a utilizar un modelo regresión lineal simple para modelar los ingresos en función del tiempo, ignorando la correlación entre medidas.

En este ejemplo, la variable dependiente Y es el ingreso anual de cada persona; mientras que la variable independiente X es el año, representando el tiempo.

El objetivo del modelo es analizar si existe una tendencia en la evolución de los ingresos y, si la hubiese, estimar la relación entre el año y el nivel de ingresos de los individuos. Sin embargo, al aplicar un modelo de regresión lineal simple, ignoraremos la dependencia entre las observaciones de cada persona, lo que resultará en una estimación sesgada y poco fiable.

La Figura 2.2 muestra la evolución de los ingresos anuales para diferentes personas a lo largo del tiempo, en el que cada línea representa a una persona. En este caso, todas las observaciones fueron tomadas en intervalos regulares (por años), lo que corresponde a la estructura de

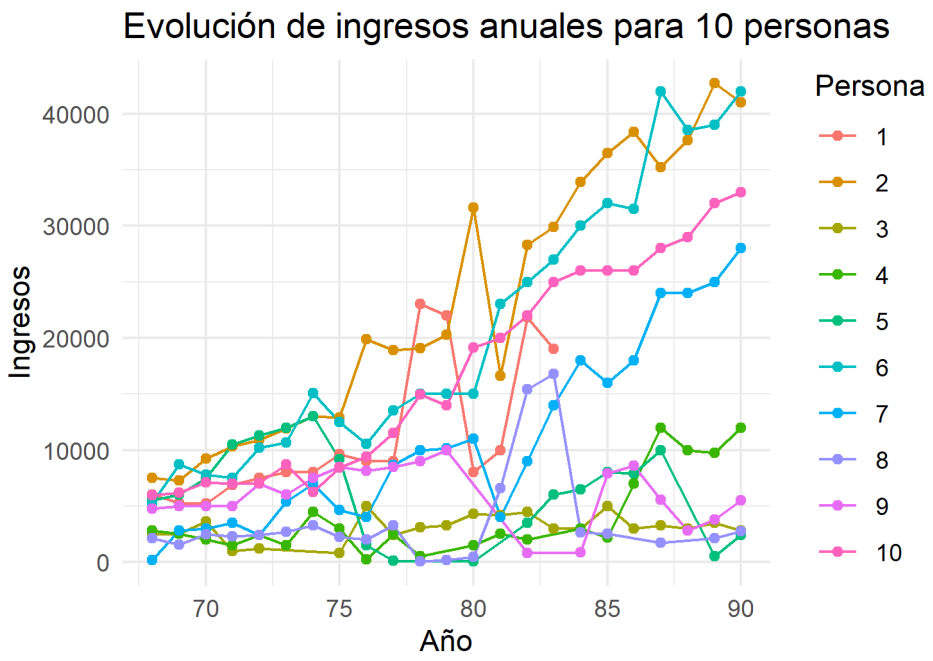

Figura 2.2: Evolución de los ingresos anuales de 10 personas a lo largo del tiempo.

medidas repetidas regulares descrita en la Figura 2.1. No obstante, se puede observar que algunos individuos presentan datos faltantes en ciertos años, lo que da lugar a una estructura con medidas regulares pero incompletas.

Esto permite mostrar cómo los ingresos varían entre individuos y años, observando que los datos son heterogéneos y varían significativamente entre individuos. Sin embargo, dentro de cada individuo, los ingresos en un año determinado tienden a ser similares a los del año anterior y el siguiente, lo que sugiere una correlación temporal en las mediciones. Además, se observa en general una tendencia creciente, aunque heterogénea, en la evolución de los ingresos, lo que refuerza la idea de una estructura dependiente en el tiempo. Esta dependencia entre observaciones dentro de cada individuo es una característica fundamental de los datos longitudinales, ya que implica que el valor de la variable en un momento dado está influenciado por valores previos del mismo individuo; algo que viola los supuestos clásicos de independencia entre observaciones.

Visto esto, modelaremos la relación entre los ingresos y el tiempo utilizando una regresión lineal simple, ignorando la dependencia entre observaciones, para mostrar las consecuencias de no cumplir las hipótesis requeridas. La Figura 2.3 muestra el ajuste de la regresión lineal simple aplicada a los datos.

La Figura 2.3 muestra cómo la regresión lineal simple aplicada a estos datos genera una representación alterada, ignorando por completo la correlación de los datos longitudinales;

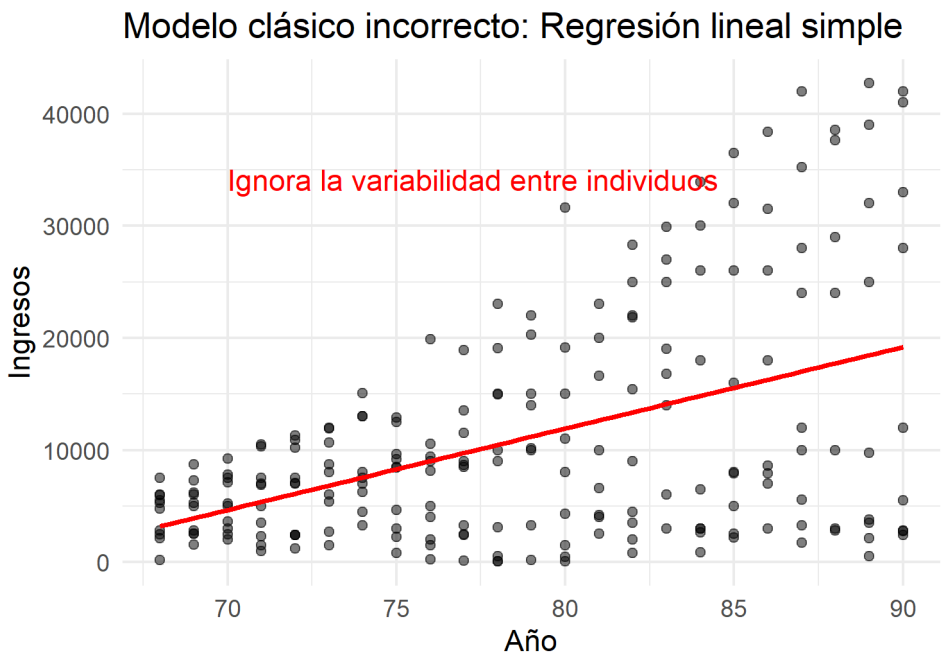

Figura 2.3: Ajuste del modelo de regresión lineal simple ignorando estructura longitudinal.

dando lugar a un mal ajuste y a resultados estadísticos no apropiados que demuestran por qué no debemos utilizar estadística clásica para este tipo de datos. No obstante, vamos a analizar la adecuación y diagnóstico del modelo para ver en detalle los motivos por los que las técnicas de estadística clásica no son las correctas para trabajar con datos longitudinales.

Al utilizar un modelo de regresión lineal simple estamos asumiendo que la variabilidad entre individuos se puede representar con un único coeficiente, ignorando por completo la dependencia entre observaciones. Para evaluar la adecuación del modelo, nos fijamos en una medida de bondad de ajuste como el coficiente de determinación, R^2. El R^2 obtenido (**0.217**) es muy bajo, indicando que el modelo explica muy poca variabilidad en los datos (21%) y que, por tanto, no nos sirve para analizar datos longitudinales ya que no captura adecuadamente la relación entre las variables.

Para realizar el diagnóstico del modelo haremos un análisis de los residuos. Recordemos que dicho análisis se basa en 4 partes fundamentales: la normalidad de los residuos, que estos tengan media cero, la no correlación de las observaciones y la homocedasticidad.

Primero de todo, vamos a analizar el supuesto de **media cero** de los residuos. Su hipótesis de asunción es la siguiente:

$$\begin{cases} H_0 : \text{Los residuos tienen una media esperada de 0.} \\ H_1 : \text{Los residuos no tienen una media esperada de 0.} \end{cases}$$

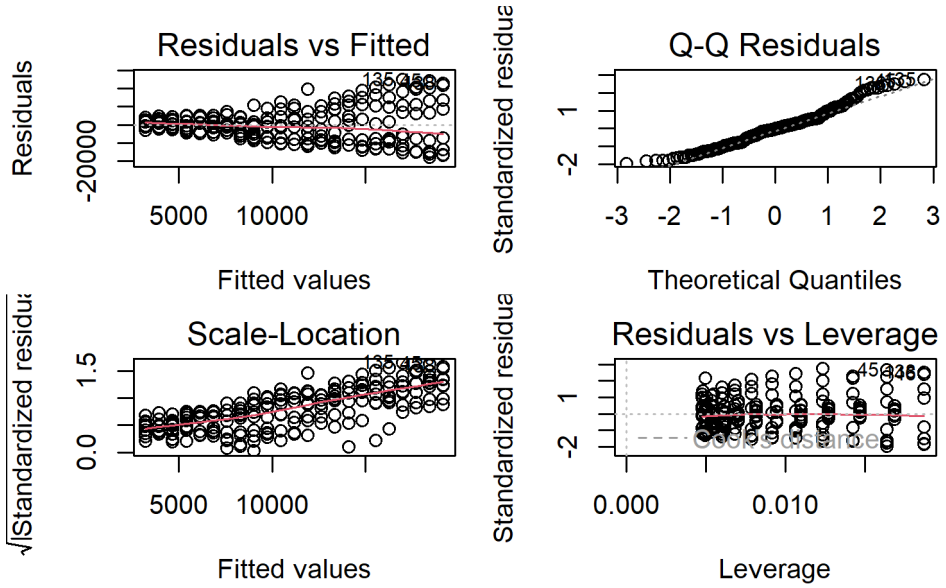

Figura 2.4: Gráfica de los residuos del modelo.

Si calculamos la media de los residuos del modelo, comprobamos que la media es **0**, pero esta no es una forma correcta de analizar la media cero ya que esto no significa que la suposición de media cero se cumpla en todas las partes del rango de valores ajustados. Para hacer un correcto análisis, nos vamos a fijar en la primera gráfica de la Figura 2.4: Residuals vs Fitted. Teóricamente, los residuos del modelo deben tener media cero, lo que implica que deberían de estar centrados sobre la línea horizontal $y = 0$. Además, si se cumple el supuesto de homocedasticidad, estos residuos deberían presentar una dispersión aproximadamente constante a lo largo del rango de valores de la variable independiente. Viendo la gráfica, podemos observar que los errores no tienen media cero ya que para los valores ajustados más altos se alejan mucho de la recta $y = 0$; por lo que esta es otra muestra más de que el modelo no es correcto para este tipo de datos.

Lo segundo que vamos a analizar es la **no correlación** entre los errores, la cual se puede analizar en la primera gráfica. Si nos fijamos en la gráfica Residuals vs Fitted, se observa un patrón curvilíneo a medida que aumenta el valor de los datos ajustados, por lo que se podría concluir que los errores están correlacionados. No obstante, para una verificación numérica haremos un test de Durbin-Watson para comprobar la no correlación. El test de Durbin-Watson verifica si los residuos están correlacionados en el tiempo. Sus hipótesis son las siguientes:

$$\begin{cases} H_0 : \text{No hay autocorrelación entre los residuos.} \\ H_1 : \text{Existe autocorrelación entre los residuos.} \end{cases}$$

En efecto, haciendo el test de Durbin-Watson vemos como el p-valor (**<0.001**) es extremadamente bajo y nos permite concluir que podemos rechazar la hipótesis nula.

Por tanto, podemos asumir que la correlación entre los errores no es 0; otro motivo más para ver que este modelo no funciona bien con datos longitudinales.

La tercera parte que vamos a analizar es la **normalidad** de los residuos. Para ello, nos fijamos en la gráfica superior derecha (Normal Q-Q) de la Figura 2.4, en la cual vemos que, aunque la mayoría de los puntos se alinean con la línea teórica, no son pocas las desviaciones que hay en los extremos; lo que sugiere que los residuos no son perfectamente normales. Para salir de dudas, podemos aplicar un test de Jarque Bera. El test de Jarque Bera comprueba si los residuos siguen una distribución normal evaluando su asimetría y curtosis. Sus hipótesis son las siguientes:

$$\begin{cases} H_0 : \text{Los residuos siguen una distribución normal.} \\ H_1 : \text{Los residuos no siguen una distribución normal.} \end{cases}$$

A través de este test, el p-valor (**0.024**) nos permite concluir que podemos rechazar la hipótesis nula y que, por tanto, los residuos no tienen normalidad.

Por último, analizaremos la **homocedasticidad** de los errores. Para ello, nos fijaremos en la primera (Residuals vs Fitted) y en la tercera gráfica (Scale-Location) de la Figura 2.4. A través de la gráfica Residuals vs Fitted, vemos como los residuos no tienen una varianza constante, sino que a medida que aumenta el valor de los valores ajustados aumenta su dispersión; por lo que no tienen homocedasticidad, sino heterocedasticidad. Mirando la gráfica Scale-Location, podemos observar una tendencia creciente por parte de los residuos que nos permite ver cómo no tienen varianza constante. Para confirmarlo, haremos un test de Breusch-Pagan. El test de Breusch-Pagan evalúa si los residuos presentan heterocedasticidad; es decir, si su varianza no es constante. Sus hipótesis son las siguientes:

$$\begin{cases} H_0 : \text{Los residuos tienen varianza constante (homocedasticidad).} \\ H_1 : \text{Los residuos no tienen varianza constante (heterocedasticidad).} \end{cases}$$

De nuevo, vemos cómo el p-valor (**<0.001**) es extremadamente pequeño, lo que nos permite rechazar la hipótesis nula y, por lo tanto, concluir que los residuos no tienen varianza constante.

A través de este análisis, hemos podido comprobar que no podemos usar modelos de estadística clásica, tal y como la regresión lineal simple, para trabajar con datos longitudinales.

Una visión más acertada sería utilizar un modelo que se ajuste a cada individuo, como se hace en la Figura 2.5.

En esta Figura 2.5, podemos observar que cada individuo se comporta de manera diferente en cuanto a la evolución de sus ingresos a lo largo del tiempo. Los interceptos y las pendientes varían considerablemente entre las personas, lo que evidencia que un único modelo no puede capturar adecuadamente la relación entre el tiempo y los ingresos para todos los

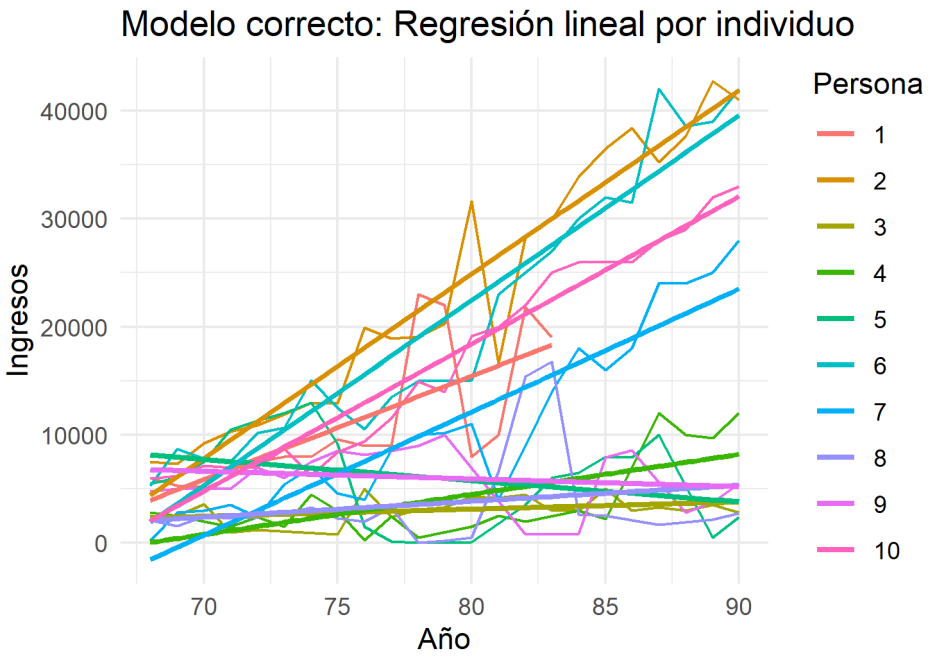

Figura 2.5: Ajuste de un modelo lineal individualizado para cada sujeto.

individuos. Este resultado destaca la heterogeneidad existente en los datos y la necesidad de utilizar modelos que consideren esta variabilidad. Al ajustar un modelo por cada individuo, capturamos mejor las características específicas de cada sujeto, pero esta estrategia presenta limitaciones: aunque mejora la representación de la variabilidad entre individuos, no permite sacar conclusiones generales sobre la población; además de que si contamos con un gran número de individuos esta aproximación no es práctica. Por ello, los **modelos mixtos**, que se explicarán en el siguiente capítulo, aparecen como una solución apropiada, ya que combinan los llamados efectos fijos y aleatorios para capturar tanto las tendencias generales de la población como las diferencias específicas entre individuos. Estos modelos ofrecen un equilibrio entre flexibilidad y generalización, respetando las características de los datos longitudinales.

3 Modelos mixtos

En este capítulo, exploraremos los **Modelos Lineales Mixtos (LMM)**, los **Modelos Lineales Generalizados (GLM)** y los **Modelos Lineales Generalizados Mixtos (GLMM)**, tres métodos estadísticos fundamentales para el análisis de datos longitudinales. Veremos cómo los LMM permiten modelar la variabilidad entre individuos a través de efectos aleatorios y fijos, facilitando el estudio de la correlación entre medidas repetidas. A continuación, introduciremos los GLM, que extienden la regresión lineal para manejar variables respuesta que no siguen una distribución normal, utilizando funciones de enlace y distribuciones de la familia exponencial. Finalmente, presentaremos los GLMM, que combinan las fortalezas de los modelos mixtos y los GLM para analizar datos longitudinales que no siguen una distribución normal incorporando tanto efectos aleatorios como funciones de enlace apropiadas. A lo largo del capítulo, revisaremos sus formulaciones matemáticas, sus hipótesis clave y cómo validarlas en la práctica.

3.1 Comparación de modelos con efectos fijos, aleatorios y mixtos

Para ilustrar estos modelos, comenzaremos con un ejemplo práctico aplicado al conjunto de datos Orthodont del paquete nlme de R (Pinheiro, Bates, y Team 2024), donde analizaremos la evolución de la distancia entre los dientes (`distance`) en función de la edad (`age`) en diferentes sujetos. Sus variables principales son:

- `distance`: distancia entre los dientes (variable respuesta).
- `age`: edad del niño (variable predictora principal).
- `Subject`: identificador del niño (variable de agrupación para efectos aleatorios).

En las siguientes secciones, compararemos tres enfoques distintos:

- Modelo con sólo **efectos fijos**: se asume que todos los sujetos siguen la misma relación.
- Modelo mixto con **intercepto aleatorio**: este modelo incluye un intercepto específico para cada individuo, permitiendo que cada sujeto tenga un valor inicial propio. Sin embargo, la pendiente que define la evolución temporal es la misma para todos los individuos. Aquí, como ya incorporamos efectos aleatorios, hablamos de un modelo lineal mixto.

- Modelo mixto con **intercepto y pendiente aleatoria**: es un modelo que incluye tanto efectos fijos como efectos aleatorios. Los efectos fijos representan patrones y tendencias globales comunes a toda la población, mientras que los efectos aleatorios modelan la variabilidad individual. En este tipo de modelos podemos introducir interceptos aleatorios, pendientes aleatorias o ambos, dependiendo de la estructura de los datos. Este método es muy útil en datos longitudinales, donde las observaciones están agrupadas por individuo y tienen una dependencia entre ellas.

3.1.1 Modelo con efectos fijos

El primer modelo que consideramos es una regresión lineal simple, en la que asumimos que la distancia interdental (`distance`) varía en función de la edad (`age`), asumiendo que todas las observaciones son independientes e ignorando su estructura jerárquica (medidas repetidas por individuo). La ecuación del modelo es:

$$distance_i = \beta_0 + \beta_1 age_i + \epsilon_i.$$

Aquí, $distance_i$ es la distancia interdental de la observación i, donde i va desde 1 hasta n, con n siendo el tamaño total de la muestra; β_0 es el intercepto común a todos los sujetos, β_1 es la pendiente (cómo cambia la distancia con la edad), y ϵ_i es el error aleatorio.

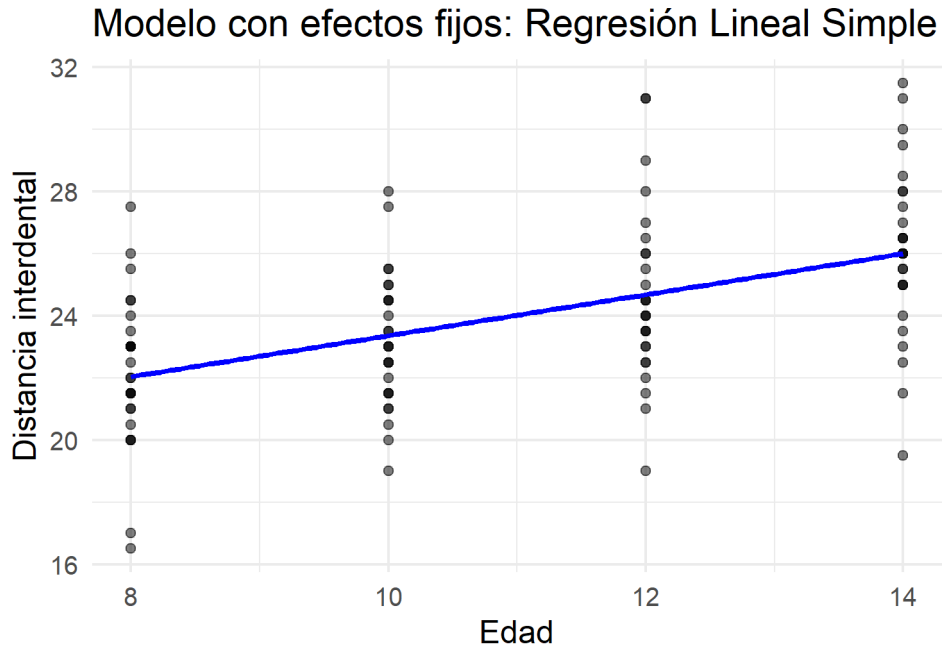

Figura 3.1: Modelo con efectos fijos: Regresión Lineal Simple.

Este modelo de la Figura 3.1 considera únicamente la edad (`age`) como predictor de la distancia (`distance`) y, como bien mencionamos antes, no tiene en cuenta que los datos son medidas repetidas de los mismos individuos, lo que puede llevar a errores de estimación debido a que se ignora por completo la correlación entre observaciones de un mismo sujeto. Como podemos comprobar a través de este ejemplo, si se ignora la estructura jerárquica, produciríamos estimaciones erróneas de la variabilidad en la población, obteniendo un coficiente de determinación R^2 muy bajo (**0.256**).

3.1.2 Modelo con intercepto aleatorio

Ahora ajustamos un modelo con efectos aleatorios, en el que permitimos que cada niño tenga su propio intercepto aleatorio (u_i), capturando de esta forma la variabilidad entre individuos. La ecuación del modelo es:

$$distance_{ij} = \beta_0 + u_i + \beta_1 age_{ij} + \epsilon_{ij}.$$

donde:

- $distance_{ij}$ representa la distancia observada para el niño i en su j-ésima medida.
- age_{ij} es la edad correspondiente a esa misma observación.
- β_0 es el intercepto poblacional, común a todos los individuos.
- u_i es el efecto aleatorio del sujeto i, que permite que cada niño tenga un intercepto diferente.
- β_1 es la pendiente común, que modela el cambio medio de la distancia con respecto a la edad.
- ϵ_{ij} es el error aleatorio, que representa la variabilidad residual no explicada por el modelo.

El subíndice i recorre los individuos (niños), mientras que j recorre las diferentes observaciones para cada individuo. De esta forma, $j = 1, \dots, n_i$, siendo n_i el número de medidas realizadas para el niño i.

Como podemos apreciar en la Figura 3.2, ahora tenemos un término que asocia a cada individuo (`Subject`) su propio intercepto aleatorio, permitiendo que la relación entre la distancia y la edad varíe entre individuos en lugar de asumir un único intercepto para todos como hacíamos en la Figura 3.1. Esto significa que algunos sujetos pueden tener distintos valores iniciales de `distance` sin afectar a la tendencia general de la población. La principal diferencia de los efectos aleatorios la podemos apreciar en la variabilidad del modelo, ya que tenemos una varianza del intercepto por sujeto de **4.472** y una varianza residual de **2.049**, lo que significa que aunque cada sujeto tiene un valor inicial diferente, todavía hay una parte de la variabilidad del modelo que no se explica por los efectos fijos ni por las diferencias entre sujetos.

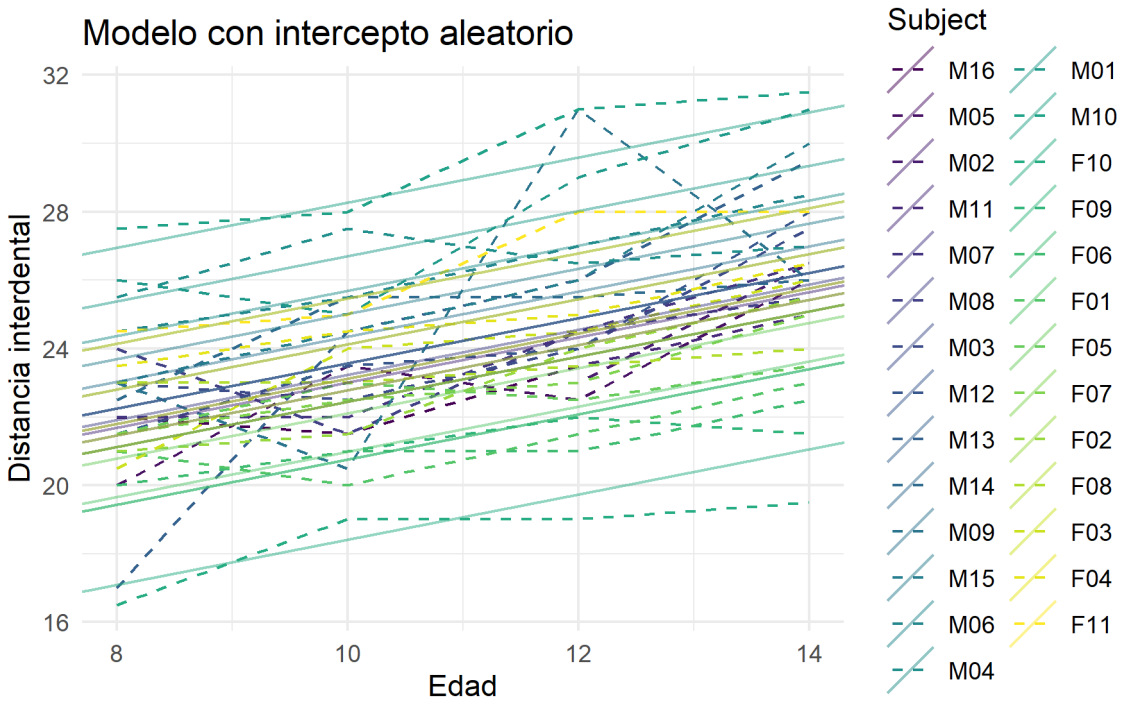

Figura 3.2: Modelo con efectos aleatorios a través de intercepto aleatorio.

Este modelo, al incorporar un intercepto aleatorio específico para cada niño, permite capturar la variabilidad individual no explicada por la edad.

3.1.3 Modelo mixto

Finalmente, ajustamos un **Modelo Lineal Mixto (LMM)** en el que consideramos tanto efectos fijos como aleatorios. Permitimos que cada niño tenga su propio intercepto (u_i) y pendiente (v_i) aleatorios, permitiendo que la relación entre edad y distancia interdental varíe entre individuos. La ecuación del modelo es:

$$distance_{ij} = \beta_0 + u_i + (\beta_1 + v_i) age_{ij} + \epsilon_{ij}.$$

Aquí u_i es el intercepto específico de cada sujeto, y v_i permite que la pendiente también varíe por individuo.

Ahora, al contar con interceptos aleatorios y permitir que la pendiente varíe entre sujetos, cada sujeto puede tener una tasa de crecimiento diferente en la distancia dental a lo largo del tiempo. Observando el modelo de la Figura 3.3, vemos que no sólo los interceptos cambian, sino que las pendientes son ligeramente distintas. Además, hemos reducido la varianza residual a **1.716**, y ahora contamos con una varianza del intercepto por sujeto de **5.417** y una variación de la

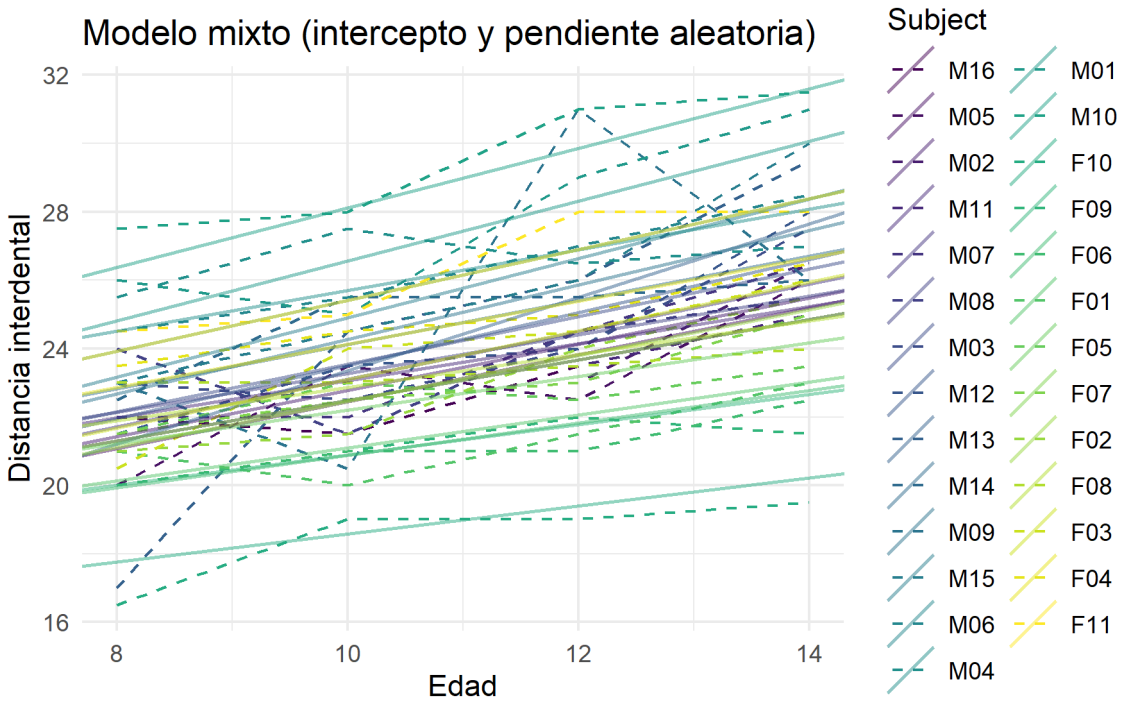

Figura 3.3: Modelo Lineal Mixto (LMM con efectos fijos y aleatorios).

pendiente entre sujetos de **0.051**. Este tipo de modelos es más realista cuando hay variabilidad individual en la evolución de la variable respuesta, además de ser más flexible ya que permite que tanto el intercepto como la pendiente varíen entre individuos.

Este último modelo generaliza la idea que vimos en el capítulo anterior, donde ajustábamos una regresión por individuo (Figura 2.5). En ese caso, teníamos una pendiente e intercepto diferentes por persona, pero ajustados de forma separada. Los modelos mixtos permiten hacer esto mismo pero de forma conjunta y efectiva, combinando la información de todos los individuos para obtener mejores estimaciones sin tener que ajustar un modelo por separado para cada uno.

Si comparamos los 3 modelos, podemos observar que el modelo con solo efectos fijos asume una única relación entre edad y distancia interdental, ignorando la variabilidad entre individuos. El modelo con intercepto aleatorio permite que cada sujeto tenga su propio intercepto, pero mantiene una pendiente común para todos. El modelo mixto (LMM) final es el más completo, permitiendo que tanto el intercepto como la pendiente varíen entre individuos. Esto demuestra la importancia de los Modelos Lineales Mixtos en el análisis de datos longitudinales, ya que incorporan tanto la variabilidad individual como la estructura jerárquica de los datos.

Antes de hablar sobre los Modelos Lineales Mixtos (LMM), conviene recordar que, en el ejemplo anterior, analizamos cómo una variable dependiente (la distancia interdental) se explicaba

a partir de una única variable independiente (la edad). Sin embargo, tanto en la regresión lineal clásica como en los modelos lineales mixtos, es posible incorporar múltiples variables explicativas que ayuden a modelar mejor la respuesta. A continuación, veremos cómo los LMM permiten esta flexibilidad, además de captar la variabilidad entre individuos a través de efectos aleatorios.

3.2 Modelos Lineales Mixtos (LMM)

Los Modelos Lineales Mixtos (LMM) son métodos estadísticos que permiten analizar datos longitudinales cuando la variable respuesta sigue una distribución normal. Uno de sus aspectos más característicos, según Francisco Hernández-Barrera en su libro *Modelos mixtos con R* (Hernández-Barrera 2024), es que asumen una relación directa entre el vector de observaciones y las covariables. Esta técnica resulta especialmente eficaz ya que permite introducir efectos aleatorios y determinar la estructura de correlación entre medidas repetidas del mismo sujeto. Además, estos modelos son robustos frente a la presencia de datos faltantes, convirtiéndolos en un método muy flexible. Una de sus principales ventajas es que permiten modelar tanto la variabilidad entre individuos, a través de efectos aleatorios, como la correlación de sus observaciones; incluyendo covariables tanto a nivel individual como grupal, y respetando la dependencia entre observaciones.

Según Julian Faraway en *Extending the Linear Model with R* (Faraway 2006), un efecto fijo es una constante desconocida que intentamos estimar a partir de los datos, mientras que un efecto aleatorio es una variable aleatoria que refleja variación individual no explicada por covariables; una distinción fundamental para interpretar correctamente los modelos. Otra de las ventajas de este tipo de modelos es su capacidad para generalizar estructuras de datos complejas, lo que hace recomendable su uso con datos longitudinales. No obstante, debemos tener en cuenta que el número de efectos aleatorios que se pueden incluir está limitado: no deben superar el número de observaciones por individuo, una restricción importante a la hora de ajustar modelos reales. En resumen, los LMM permiten modelar al mismo tiempo los efectos comunes a toda la población (efectos fijos) y la variabilidad individual (efectos aleatorios), proporcionando estimaciones más precisas y respetando la estructura de dependencia de los datos longitudinales.

La ecuación para este tipo de modelos, en los que y_{ij} representa la observación j-ésima del individuo i:

$$y_{ij} = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ijk} + u_{0i} + \sum_{k=1}^{K} u_{ki} x_{ijk} + e_{ij}.$$

- $i = 1, \dots, N$: indica el individuo o sujeto.
- $j = 1, \dots, n_i$: indica la observación o medición j-ésima del individuo i. Cada sujeto puede tener un número diferente de observaciones.

- $k = 1, \ldots, K$: indica las variables explicativas o predictoras incluidas en el modelo, siendo K el número total de variables independientes.
- β_0 y β_k son los efectos fijos (intercepto y pendientes comunes a todos los individuos). En conjunto, modelan el efecto medio de los predictores sobre la respuesta.
- u_{0i} y u_{ki} son los efectos aleatorios (variaciones individuales del intercepto y de las pendientes), los cuales se asumen que siguen una distribución normal de media cero. Cada individuo i puede tener una pendiente propia para cada predictor x_{ijk}, lo que permite capturar variaciones individuales en la relación entre las variables explicativas y la variable respuesta.
- x_{ijk} son las variables explicativas.
- e_{ij} es el término de error residual, que recoge la variación no explicada por el modelo.

En la práctica, este modelo se especifica en R mediante la función `lmer()` del paquete `lme4` (Bates et al. 2015).

Una vez formulado el modelo, necesitamos estimar los parámetros, contando con dos métodos principales de estimación como bien se explica en *Modelos mixtos con R* (Hernández-Barrera 2024). El primero es el método de máxima verosimilitud (ML), que utiliza la función de verosimilitud completa del modelo. Este método estima tanto los efectos fijos como las varianzas de los efectos aleatorios, y es útil para comparar modelos con diferentes efectos fijos. El otro método es el de máxima verosimilitud restringida (REML), que estima solo las varianzas de los efectos aleatorios, ajustando los grados de libertad para evitar sesgo en la estimación de la varianza. El REML es el método preferido para comparar modelos con la misma estructura de efectos fijos, pero distinta estructura de efectos aleatorios.

Cuando ya se ha ajustado el modelo, el siguiente paso es seleccionar la estructura más adecuada de efectos fijos y aleatorios. La selección del modelo es clave en el análisis de datos longitudinales, ya que el hecho de incluir o descartar ciertos efectos puede afectar notablemente la calidad del ajuste y la interpretación. En la práctica, se recomienda seguir una estrategia jerárquica, comenzando por un modelo completo que incluya todas las variables candidatas como efectos fijos, así como una estructura lo más completa posible de efectos aleatorios. Para comparar modelos con distinta estructura de efectos fijos, debe utilizarse el método de máxima verosimilitud (ML), debido a que la función de verosimilitud incorpora estos efectos y permite su comparación. Una vez seleccionada la mejor combinación de efectos fijos, se recomienda ajustar el modelo final con máxima verosimilitud restringida (REML), que proporciona estimaciones más precisas de los componentes de varianza (efectos aleatorios).

Además de comparar modelos mediante ML o REML, se utilizan criterios de información como el AIC (Akaike Information Criterion) o el BIC (Bayesian Information Criterion). Ambos equilibran el ajuste del modelo con su complejidad, llegando a penalizar la inclusión de demasiados parámetros. En particular, el BIC tiende a ser más conservador que el AIC al penalizar más los modelos complejos.

Por último, es esencial validar el modelo ajustado. Esto se realiza evaluando las asunciones sobre los residuos del modelo, tal como se hace en la regresión clásica. En concreto, un gráfico de residuos estandarizados vs valores predichos debe mostrar una nube de puntos sin estructura aparente, lo que indica homocedasticidad; y en el QQ-plot, que permite evaluar la normalidad de los residuos, los puntos deben seguir aproximadamente la diagonal si queremos asumir normalidad. Aunque algunos métodos para validar modelos mixtos utilizan los Empirical Bayes Estimates (EBEs) para evaluar la distribución de los efectos aleatorios, en este trabajo optamos por utilizar el paquete `DHARMa` (Hartig 2024), que ofrece una validación más sistemática de los supuestos del modelo a través de simulaciones de residuos. A diferencia de los EBEs, que pueden inducir sesgos al tratarse de estimaciones condicionadas, `DHARMa` genera residuos simulados independientes de la estructura del modelo, permitiendo detectar problemas como falta de normalidad, heterocedasticidad o dependencia estructural. Este método resulta especialmente apropiado cuando se trabaja con estructuras de datos complejas, como en nuestro caso con datos longitudinales.

3.3 Modelos Lineales Generalizados (GLM)

En la sección anterior trabajamos con modelos lineales mixtos (LMM), los cuales asumen que la variable respuesta sigue una distribución normal, que la relación entre los predictores y la respuesta es lineal y que los residuos presentan varianza constante. Sin embargo, en muchas situaciones reales estas asunciones no se cumplen, ya que la variable de interés puede ser binaria o puede no modelarse adecuadamente con una distribución normal. Ante este tipo de limitaciones, resulta necesario generalizar el modelo para poder adaptarlo a otras distribuciones y a relaciones no lineales entre predictores y respuesta. Para ello, introduciremos los Modelos Lineales Generalizados (GLM), que permiten modelar variables respuesta de distintas familias de distribuciones manteniendo una estructura lineal en los predictores, a través de funciones de enlace. Los Modelos Lineales Generalizados resultan clave para desarrollar modelos mixtos aún más generales y aplicables a una mayor variedad de contextos.

Los Modelos Lineales Generalizados son una generalización de los modelos lineales para una variable respuesta perteneciente a la familia exponencial, en la que tenemos una función de enlace que describe cómo la media de la variable respuesta y la combinación lineal de variables explicativas están relacionadas. Los GLM son una clase de modelos más amplia que tienen formas parecidas para sus varianzas y verosimilitudes, generalizando la regresión lineal múltiple (Roback y Legler 2021).

La familia exponencial tiene esta forma:

$$f(y \mid \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right].$$

En esta ecuación, θ es el **parámetro canónico** y representa la posición; mientras que ϕ es el **parámetro de dispersión** y representa la escala. De la misma forma, a, b y c representan

diferentes miembros de la familia exponencial. En función del parámetro de dispersión, podemos distinguir entre familias exponenciales de un parámetro, y familias exponenciales de dos parámetros.

Como familias exponenciales de un parámetro, tenemos las distribuciones de Poisson y la Binomial. Vamos a demostrar que la distribución de Poisson es, en efecto, una familia exponencial de un parámetro.

Para ello, aplicando propiedades logarítmicas, podemos definir la distribución de Poisson como:

$$P(Y = y) = e^{-\lambda} e^{y \log \lambda} e^{-\log(y!)} = e^{y \log \lambda - \lambda - \log(y!)}.$$

Si comparamos esta función de masa de probabilidad con la función de probabilidad general para familias con un único parámetro, podemos ver que:

$$\begin{aligned} a(y) &= y \\ b(\theta) &= \log(\lambda) \\ c(\theta) &= -\lambda \\ d(y) &= -\log(y!) \end{aligned}$$

La función $b(\theta)$ es lo que denominamos **enlace canónico**, una función que nos permite modelar como una función lineal de variables explicativas.

Como familias exponenciales de dos parámetros, tenemos la distribución gamma y la normal. De forma parecida a la anterior, demostraremos que la distribución normal es una familia exponencial de dos parámetros.

Podemos definir la función de densidad de una distribución normal como:

$$f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right).$$

Si separamos términos y los escribimos como términos logarítmicos, tenemos que:

$$f(y|\mu, \sigma^2) = \exp\left(y \cdot \frac{\mu}{\sigma^2} - \frac{y^2}{2\sigma^2} + \left(-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)\right).$$

Si comparamos esta función de densidad con la forma general de la familia exponencial, podemos ver que:

$$\begin{aligned} a(y) &= y \\ b(\mu, \sigma^2) &= \frac{\mu}{\sigma^2} \\ c(\mu, \sigma^2) &= -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \\ d(y, \sigma^2) &= -\frac{y^2}{2\sigma^2} \end{aligned}$$

Por lo tanto, demostramos que la distribución normal también pertenece a la familia exponencial, pero con una peculiaridad respecto a la distribución de Poisson: es una familia exponencial de dos parámetros, la media μ y la varianza σ^2. En este caso, el término $b(\mu, \sigma^2)$ es el enlace canónico que conecta las variables explicativas con el modelo.

En concreto, para los casos en los que la respuesta no es normal, la ecuación del modelo es la siguiente:

$$g\left(E(y_{ij} \mid x_{ijk}, \beta_{0i}, \dots, \beta_{Ki})\right) = \beta_{0i} + \sum_{k=1}^{K} \beta_{ki} x_{ijk}.$$

Donde g es la función enlace y, pese a que puede parecerse mucho a la función para modelos LMM, tienen algunas diferencias, como que en el primer miembro tenemos el enlace del valor esperado en vez de la variable respuesta, y en el segundo miembro no se cuenta con los errores; por lo que no existe una matriz de correlaciones de los residuos. De esta forma, ya hemos generalizado nuestro modelo para manejar variables respuesta que no siguen una distribución normal. A través de esta generalización, podemos escribir la función de masa o densidad de probabilidad de distintas distribuciones para poder modelar el enlace canónico como función lineal de las variables predictoras.

Una vez formulado el modelo GLM, los parámetros se estiman mediante el método de máxima verosimilitud (ML), que consiste en encontrar los valores de los coeficientes que maximizan la probabilidad de observar los datos reales. Como veremos más adelante, utilizaremos modelos GLM con funciones como `glm()` (R Core Team 2024b) donde se emplea por defecto el método de máxima verosimilitud (ML) como técnica de estimación. Este método se implementa mediante el algoritmo Iteratively Reweighted Least Squares (IRLS), a través del cual se encuentran los parámetros óptimos ajustando iterativamente los pesos de las observaciones según su varianza esperada. Esta estrategia es apropiada para distribuciones como la binomial o la gamma y permite un ajuste robusto en situaciones donde la variable respuesta no sigue una distribución normal.

En cuanto a la validación, se analizan medidas como la devianza o el AIC para comparar modelos. También es importante evaluar la distribución de la variable respuesta y el ajuste del modelo a través de gráficos de residuos.

3.3.1 Ejemplo práctico

Supongamos que queremos modelar el número de llamadas que recibe un centro de emergencias por hora (`llamadas`), en función del número de operadores de guardia (`operadores`). Este tipo de datos es típico de una distribución Poisson, por lo que analizar estos datos mediante una regresión lineal tradicional no sería apropiado.

En este modelo, estamos modelando el logaritmo del número esperado de llamadas como una función lineal del número de operadores. Tras ajustar el modelo a una muestra simulada de

100 observaciones, obtenemos la siguiente expresión para el número esperado de llamadas por hora:

$$\mathbb{E}[\text{llamadas}] = \exp(0.658 + 0.258 \cdot \text{operadores}).$$

La interpretación de los coeficientes es en términos del logaritmo de la tasa: por cada operador adicional, el número esperado de llamadas se multiplica por **1.294**; es decir, aumenta un 29.4%.

3.4 Modelos Lineales Generalizados Mixtos (GLMM)

Como ya hemos mencionado anteriormente, cuando trabajamos con datos longitudinales cuya variable respuesta no sigue una distribución normal, los Modelos Lineales Mixtos (LMM) dejan de ser apropiados. En estos casos, podemos extender los modelos hacia los Modelos Lineales Generalizados Mixtos (GLMM), los cuales combinan la flexibilidad de los GLM con la estructura de efectos aleatorios de los LMM. Un GLMM permite modelar variables respuesta que pertenecen a la familia exponencial (binomial, Poisson, etc.), y la correlación entre medidas repetidas para el mismo individuo mediante efectos aleatorios.

La ecuación general de un GLMM (McCulloch, Searle, y Neuhaus 2008) es:

$$g\left(\mathbb{E}(y_{ij} \mid \mathbf{b}_i)\right) = \mathbf{x}_{ij}^\top \beta + \mathbf{z}_{ij}^\top \mathbf{b}_i.$$

Donde:

- y_{ij} es la respuesta del individuo i en la ocasión j.
- $\mathbf{x}_{ij}$ es el vector de covariables con efectos fijos.
- β es el vector de coeficientes fijos.
- $\mathbf{z}_{ij}$ es el vector de covariables con efectos aleatorios.
- $\mathbf{b}_i$ es el vector de efectos aleatorios del individuo i, que se asume que sigue una distribución normal multivariante $\mathbf{b}_i \sim \mathcal{N}(0, \mathbf{D})$.
- $g(\cdot)$ es la función de enlace, que conecta la media condicional de la respuesta con la combinación lineal de predictores.

Dependiendo de la naturaleza de y_{ij}, tendremos que utilizar distintos enlaces. Para datos binarios, usaremos un enlace logit y distribución binomial; para datos de conteo, utilizaremos un enlace log y distribución de Poisson; y para tiempos o proporciones, emplearemos enlaces adaptados como log-log, logit, etc. Si la variable respuesta es binaria (por ejemplo, éxito/fracaso), se usa un modelo logístico mixto:

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij} + b_i.$$

Donde:

- $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$
- $p_{ij} = \mathbb{P}(y_{ij} = 1 \mid b_i)$
- $b_i \sim \mathcal{N}(0, \sigma_b^2)$ es un efecto aleatorio por sujeto.

Esto permite modelar probabilidades condicionales considerando la variabilidad entre individuos.

3.4.1 Ejemplo práctico

Supongamos que queremos modelar si un estudiante aprueba un examen (`aprobado = 0/1`) en función de las horas de estudio (`horas`) y si el estudiante forma parte de un grupo diferente (`grupo`). Como los estudiantes tienen múltiples exámenes, añadimos un efecto aleatorio por estudiante.

En este modelo GLMM (`glmer`) (Bates et al. 2015) estamos modelando la probabilidad de aprobar un examen en función de dos variables explicativas: horas (número de horas de estudio) y grupo (grupo educativo (A o B)). Además, incluimos un efecto aleatorio de intercepto por estudiante, lo cual es adecuado porque cada estudiante tiene múltiples observaciones (exámenes), y esperamos que haya variabilidad entre ellos.

El intercepto aleatorio por estudiante tiene una desviación estándar de **0.948**, lo que indica que hay una variación importante en la tendencia a aprobar entre estudiantes, incluso tras haber controlado mediante las horas de estudio y el grupo; lo que justifica el uso de un GLMM en lugar de un GLM clásico. Este problema justifica cómo un GLMM puede capturar variabilidad individual (entre estudiantes) y a la vez evaluar el impacto de efectos fijos. El uso del modelo mixto es clave, ya que si ignoramos el efecto aleatorio por estudiante, estaríamos asumiendo que todos los estudiantes tienen la misma tendencia a aprobar, lo cual podemos apreciar no es el caso según la varianza estimada.

Los GLMM tienen múltiples ventajas respecto a otros modelos, ya que permiten ajustar modelos a variables respuesta no continuas, incorporan variabilidad entre individuos mediante efectos aleatorios, se adaptan bien a datos longitudinales, y permiten sacar conclusiones globales considerando la dependencia temporal. Por tanto, los GLMM son un método esencial para el análisis de datos longitudinales cuando la respuesta no es normal, ya que preservan la estructura de dependencia de los datos sin violar los supuestos del modelo.

La estimación de los parámetros en modelos GLMM es más compleja que en GLM o LMM, ya que la función de verosimilitud no tiene una forma analítica cerrada. Por ello, se utilizan técnicas de aproximación numérica como la aproximación de Laplace, que integra los efectos aleatorios y aproxima la verosimilitud, la cuadratura Gauss-Hermite adaptativa, que mejora la precisión en presencia de muchos efectos aleatorios, o Penalized Quasi-Likelihood (PQL),

en versiones simplificadas. Estas aproximaciones permiten sacar conclusiones sobre los efectos fijos y estimar la variabilidad entre sujetos. Además, los GLMM pueden presentar problemas de convergencia, especialmente con estructuras complejas o tamaños muestrales pequeños. Es importante validar el modelo revisando los residuos, la bondad de ajuste (AIC, BIC), y la significancia de los efectos aleatorios, por ejemplo, mediante test de razón de verosimilitudes anidados o comparaciones de modelos.

En la práctica, utilizaremos la función `glmmTMB` (M et al. 2025) para ajustar modelos GLMM. Esta elección se debe a que `glmmTMB` ofrece una mayor flexibilidad que otras alternativas como `glmer`, permitiendo especificar una amplia variedad de distribuciones (incluidas binomial, Poisson, gamma o beta) y enlaces personalizados. Además, presenta una robustez computacional que resulta de gran utilidad en contextos con estructuras complejas de efectos aleatorios. Esta versatilidad lo convierte en una herramienta adecuada para los distintos tipos de variables respuesta que pueden encontrarse en el análisis de datos longitudinales.

3.5 Validación del modelo y predicciones

Una vez ajustado un modelo mixto, debemos comprobar que el modelo se adapta adecuadamente a los datos y que cumple los supuestos teóricos necesarios antes de realizar predicciones o extraer conclusiones. En modelos LMM, esto se produce mediante un análisis de residuos (residuos vs ajustados, QQ-plot, trayectorias individuales) que realizaremos con el paquete `DHARMa`, el cual genera residuos simulados sobre los que aplicar tests de uniformidad, dispersión y detección de valores atípicos. En modelos GLMM, aunque el principio es parecido, la validación debe adaptarse al tipo de variable respuesta, utilizando medidas como el AIC para comparar modelos y comprobaciones como la convergencia del ajuste o la significancia de los efectos aleatorios. Validar el modelo permiten concluir si el modelo es adecuado para los datos; garantizando que sus predicciones sean fiables y que las conclusiones son robustas.

Una vez validado el modelo, este puede ser utilizado para hacer predicciones. En el caso de un modelo mixto bien ajustado, las predicciones permiten estimar el valor esperado de una variable en un momento determinado considerando tanto la tendencia general como las diferencias individuales. A través del uso de modelos validados, los LMM y GLMM se postulan como métodos muy eficaces a la hora de generar estimaciones, hacer predicciones futuras o evaluar situaciones hipotéticas.

4 Análisis exploratorio de la base de datos

Como se ha señalado en la introducción, el trabajo se centra en el desarrollo práctico de un ejemplo de análisis de datos longitudinales, con el fin de aplicar las técnicas estadísticas descritas en los capítulos anteriores. En este capítulo, en concreto, se presenta la base de datos principal utilizada, junto con las fuentes complementarias, y se lleva a cabo un análisis exploratorio en profundidad. Este análisis inicial resulta clave para comprender la estructura de los datos, identificar posibles patrones, detectar valores atípicos o faltantes, y orientar las decisiones de modelización que se abordarán en los capítulos siguientes.

4.1 Análisis exploratorio inicial

El conjunto de datos World Happiness (2015-2024) recopila información sobre la felicidad percibida en diferentes países a lo largo de los años. Esta base de datos proviene de los informes anuales de felicidad publicados por la Red de Soluciones para el Desarrollo Sostenible de la ONU, los cuales se basan en encuestas realizadas a los ciudadanos nivel mundial. La base tiene una buena cobertura temporal, ya que abarca datos de 2015 a 2024 y permite analizar tendencias a lo largo del tiempo, pero también tiene una buena cobertura geográfica porque incluye información de diferentes países y regiones del mundo. Es ampliamente utilizada en estudios de bienestar y calidad de vida, y contiene métricas económicas y sociales que permiten un análisis estadístico y comparativo. Cada fila representa un país en un año determinado y contiene variables socioeconómicas y de bienestar que pueden influir en la percepción de felicidad de su población. Estas variables son:

- `Country`: nombre del país.
- `Region`: continente o agrupación geográfica del país.
- `Happiness Score`: puntuación de felicidad promedio en el país.
- `GDP per capita`: Producto Interno Bruto (PIB) per cápita. Mide la riqueza económica del país.
- `Social Support`: medida de apoyo social basada en la percepción de las personas sobre la ayuda que pueden recibir de familiares y amigos.
- `Healthy Life Expectancy`: esperanza de vida saludable en años.
- `Freedom to Make Life Choices`: libertad para tomar decisiones personales.

- `Generosity`: nivel de generosidad en la sociedad, basado en donaciones y ayuda a otros.
- `Perceptions of Corruption`: nivel de percepción de corrupción en el gobierno y los negocios.

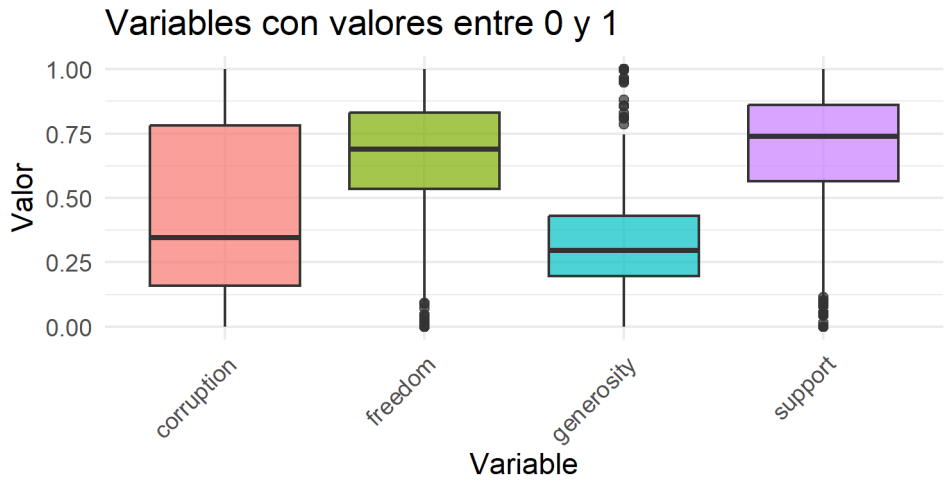

Figura 4.1: Distribución de las variables con rango entre 0 y 1 de la base de datos World Happiness.

Tanto la Figura 4.1 como Figura 4.2 nos muestran la distribución de las diferentes variables, en las que se aprecia que no hay valores atípicos. Explorando opciones para enriquecer la base de datos y, de esta manera, construir un modelo más informativo, se han estudiado diversas alternativas y se ha decidido integrar información de dos fuentes externas que aportan indicadores políticos y de libertades civiles en los países. Estas bases de datos nos permitirán explorar hasta qué punto la calidad de vida, la democracia, los derechos políticos y las libertades influyen en la percepción de felicidad de los ciudadanos.

La primera base de datos que hemos considerado es "Freedom in the World" (Freedom House 2024), un informe anual de la organización Freedom House, que evalúa el estado de los derechos políticos y libertades civiles a nivel global. Se clasifica cada país en función de indicadores de democracia, participación política y derechos individuales. El motivo por el que hemos elegido esta base de datos es porque los estudios en ciencias sociales han mostrado que la percepción de felicidad no solo está ligada a factores económicos, sino también a la capacidad de los ciudadanos para expresarse libremente, participar en política y vivir sin restricciones autoritarias. Por ejemplo, Inglehart et al. (2008) señalan que la libre elección está positivamente correlacionada con mayores niveles de felicidad. De manera similar, Helliwell et al. (2023) destacan que el apoyo institucional y la libertad individual son factores clave del Happiness Score en el World Happiness Report. Incorporar estos datos nos permitirá ver si existe dicha correlación entre los niveles de libertad y la felicidad percibida en cada país. Como esta base

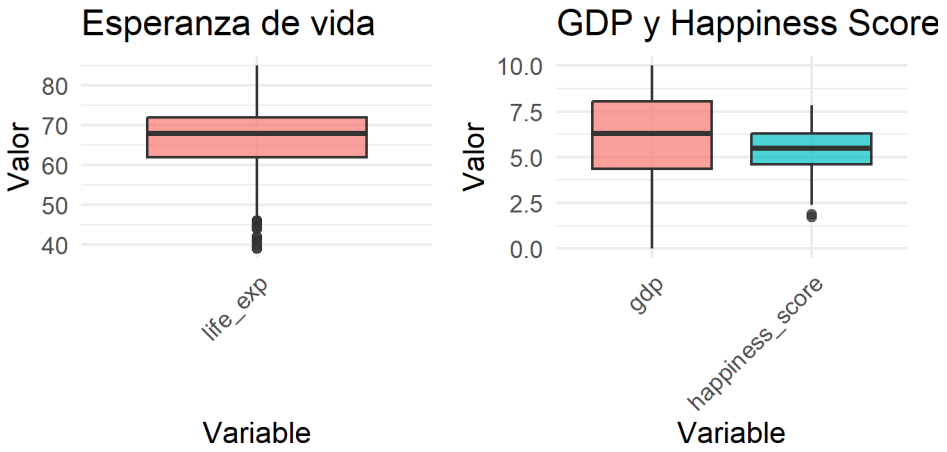

Figura 4.2: Distribución de las variables de la base de datos World Happiness.

de datos cuenta con una gran cantidad de variables, hemos elegido las siguientes variables de interés:

- `Country/Territory`: identificación del país o territorio.
- `Region`: indica la zona geográfica del país.
- `c/T`: diferencia entre países y territorios, aunque este concepto puede ser delicado según el análisis.
- `Edition`: año del reporte, fundamental para el análisis longitudinal.
- `Status`: clasificación del país en cuanto a su libertad: Libre (F), Parcialmente Libre (PF) o No Libre (NF).
- `PR rating (Political Rights)`: puntuación de 1 a 7 sobre derechos políticos.
- `CL rating (Civil Liberties)`: puntuación de 1 a 7 sobre libertades civiles.

La otra base de datos que hemos elegido para nuestro análisis es "Democracy Data", una base de datos que proviene del proyecto TidyTuesday (R4DS Online Learning Community 2024) y está basado en estudios académicos sobre democracia y regímenes políticos. Esta base de datos clasifica a los países según su sistema de gobierno y proporciona información detallada sobre su historia política. Una de las características que tiene esta base de datos es que incluye información hasta 2020, por lo que tenemos que considerar que, si vamos a trabajar con ella, tendremos las características completas de las observaciones; pero en un período reducido de tiempo. Dado que la felicidad no solo depende de factores económicos, sino también de la estabilidad política y la forma de gobierno de un país, estas variables pueden ayudarnos a explicar por qué algunos países presentan niveles bajos de felicidad a pesar de tener una economía sólida. Al igual que en el caso anterior, como esta base de datos contiene más de 40 variables, hemos decidido quedarnos con aquellas que consideramos que mejor se adaptan a nuestro análisis. Estas variables son:

- `country_name`: nombre del país.
- `year`: año de observación.
- `regime_category`: clasificación del sistema de gobierno (democracia parlamentaria, autocracia civil, dictadura militar, monarquía, etc.).
- `is_monarchy`: indica si el país es una monarquía.
- `is_democracy`: indica si el país es una democracia.
- `is_presidential`: indica si el sistema es presidencialista.
- `is_colony`: identifica si el país sigue siendo una colonia.
- `is_communist`: indica si el país sigue un régimen comunista.
- `spatial_democracy`: evalúa el nivel de democracia en los países vecinos.
- `has_full_suffrage`: indica si hay sufragio universal.
- `electoral_category`: tipo de elecciones (no democráticas, de partido único, multipartidistas no democráticas o democráticas).
- `spatial_electoral`: evalúa la calidad electoral de los países vecinos.
- `has_free_and_fair_election`: indica si las elecciones en el país son libres y justas.
- `has_alternation`: indica si existe alternancia en el poder.

Agregando estas bases de datos consideramos que podemos abarcar de forma bastante completa los diferentes factores que afectan a la percepción de la felicidad, lo que nos permitirá realizar un análisis más profundo y sacar conclusiones sobre la relación entre política, democracia y bienestar en distintos países. Para combinar adecuadamente estas fuentes, las variables clave utilizadas para relacionarlas fueron el país o región y el año, asegurando la coherencia temporal y geográfica en el análisis.

4.2 Análisis de las bases de datos complementarias

Al trabajar con datos longitudinales a nivel estatal, surge una cuestión clave: ¿en qué medida las características políticas de cada país se mantienen estables o experimentan cambios relevantes a lo largo del tiempo? Para responder a esta pregunta, incorporaremos la información de las bases de datos Freedom in the World y Democracy Data que permite capturar distintos aspectos del contexto político de los países, como el tipo de régimen, la alternancia en el poder o la existencia de elecciones libres.

Sin embargo, nos encontramos con dos limitaciones importantes. En primer lugar, muchas de estas variables presentan una cobertura temporal limitada, con registros ausentes en varios años dentro del periodo 2015–2024. En segundo lugar, se trata de variables que tienden a ser bastante estables, ya que en la mayoría de los países analizados no se observan cambios relevantes a lo largo del tiempo. Estas circunstancias suponen un obstáculo al querer tratarlas como variables longitudinales.

Por ello, incorporaremos estas variables como una "foto fija" del contexto político, extrayendo su valor más representativo dentro del periodo estudiado. De esta forma, las utilizamos como

covariables fijas en los modelos de felicidad, de manera que enriquecemos el análisis sin introducir sesgos derivados de valores faltantes o variaciones mínimas que podrían aportar ruido en lugar de información de valor.

No obstante, existen excepciones que tenemos que tener en cuenta, ya que algunos países sí que han experimentado cambios significativos en su sistema político, forma de gobierno o nivel de democracia. A continuación se representa gráficamente la evolución de dichas variables para estos casos específicos, con el objetivo de ilustrar de forma visual estos cambios.

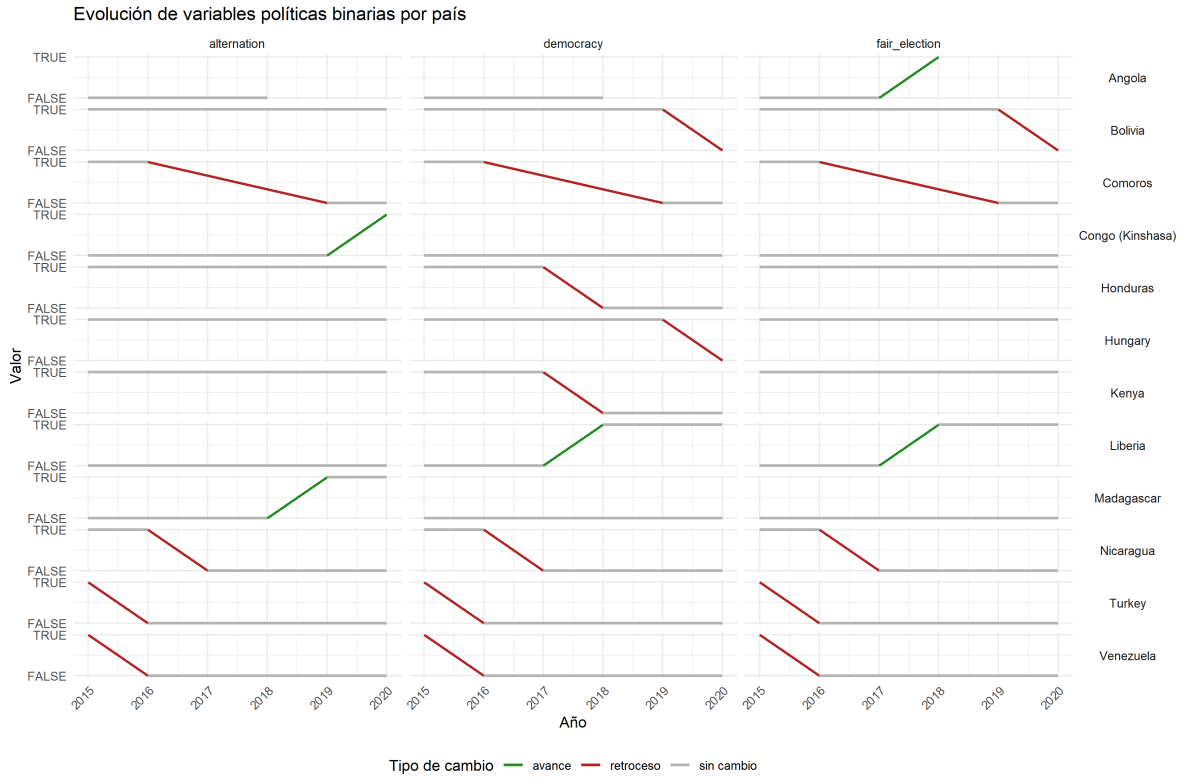

Figura 4.3: Evolución temporal de las variables políticas binarias de la base de datos Democracy Data.

Estas figuras sobre la evolución de diferentes variables políticas muestran de forma clara y visual la evolución temporal de varios países que han experimentado cambios entre 2015 y 2020.

En cuanto a los cambios en la alternancia en el poder, los cuales podemos apreciar en la Figura 4.3, vemos cómo en Nicaragua y Venezuela esta variable refleja la falta de alternancia, mostrando la acumulación del poder por parte de gobiernos autoritarios. Observando la variable democrática, vemos que países como Nicaragua y Hungría dejaron de considerarse estados democráticos una vez llegaron al poder figuras autoritarias como Daniel Ortega

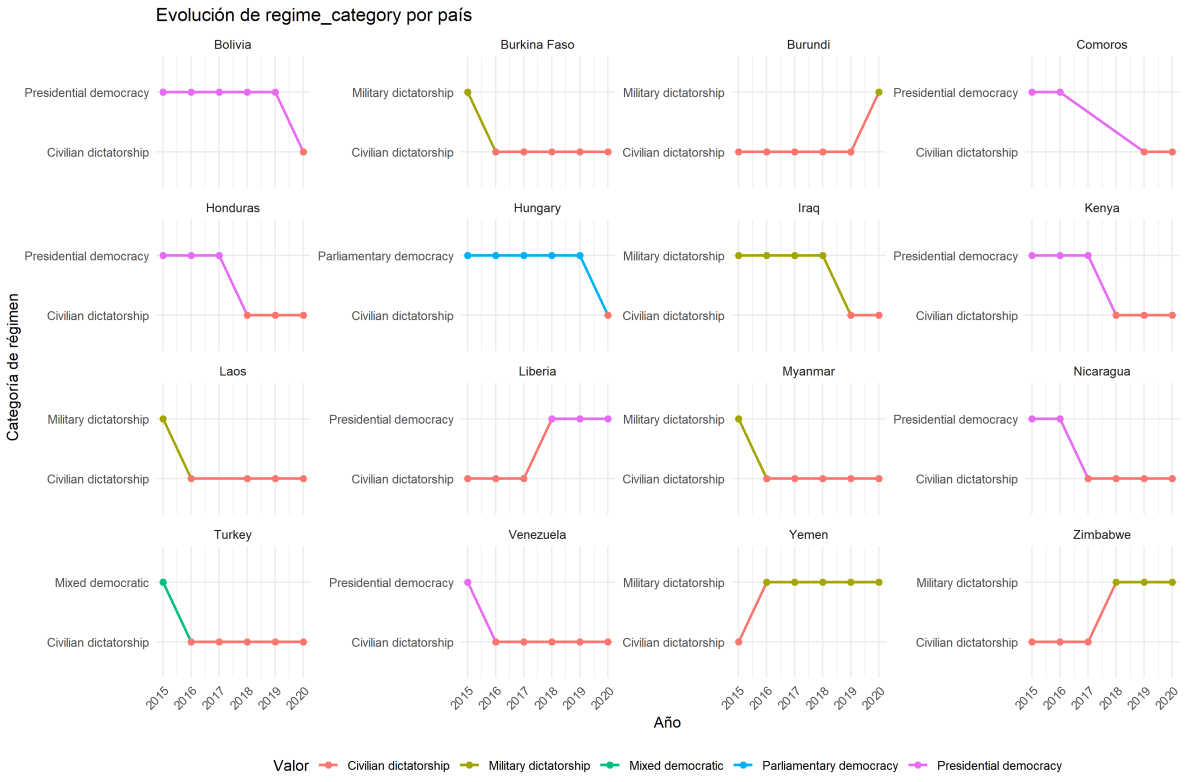

Figura 4.4: Evolución del cambio de régimen democrático en países desde 2015 a 2020.

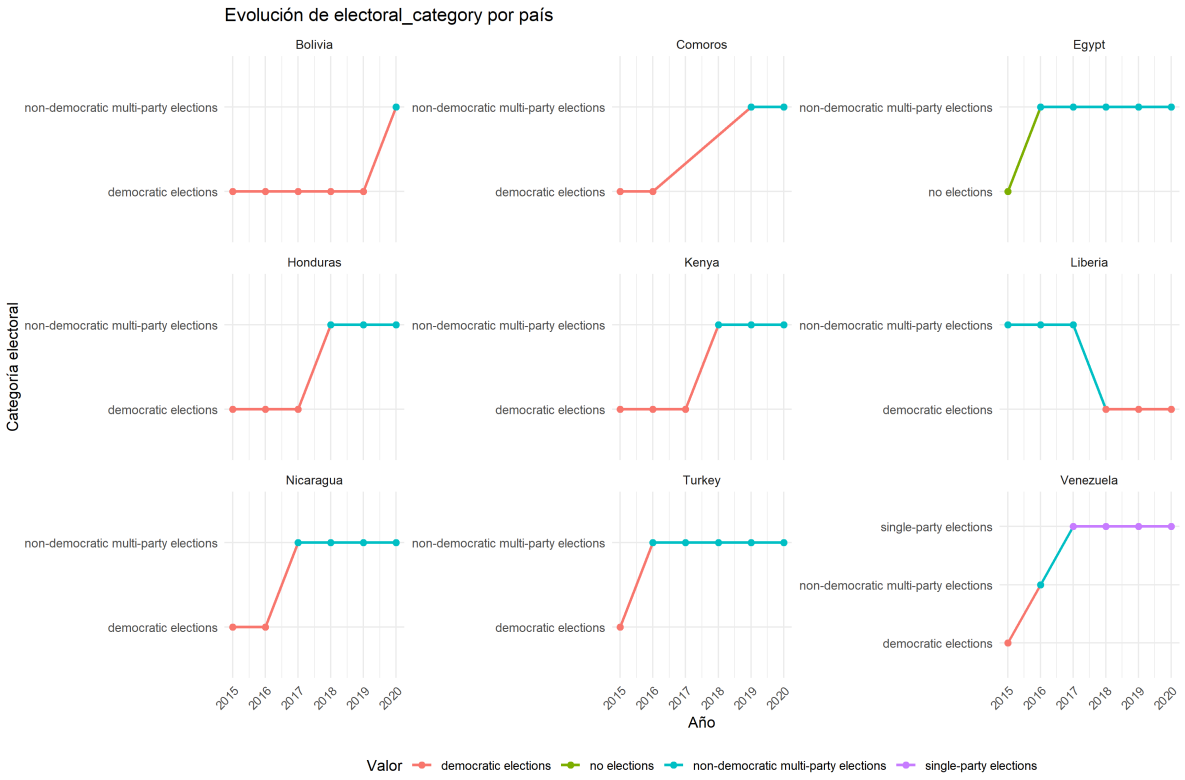

Figura 4.5: Evolución del cambio de categoría electoral en países desde 2015 a 2020.

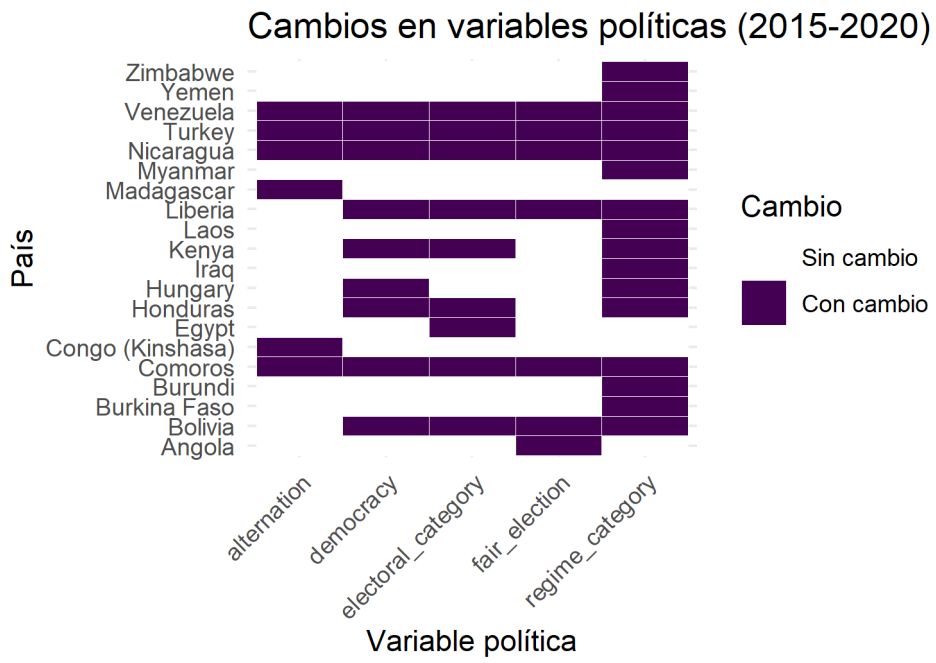

Figura 4.6: Análisis de los distintos cambios políticos en países de 2015 a 2020.

y Viktor Orbán. Por último, si nos fijamos en la variable sobre elecciones libres y justas, vemos que Bolivia presenta un cambio negativo en 2019 correspondiente a las acusaciones de fraude electoral; provocando una crisis política y la renuncia del aquel entonces presidente Evo Morales. Sin embargo, no todo son cambios negativos, ya que en la Figura 4.3 también podemos observar avances en países como Liberia que han mejorado sus instituciones democráticas gracias a la llegada al poder del exfutbolista George Weah en las elecciones de 2017; las cuales fueron históricas ya que supusieron la primera transición presidencial pacífica en el país desde 1944.

Si observamos los cambios en el tipo de régimen en la Figura 4.4, vemos como Hungría pasó de ser considerada una democracia parlamentaria a una dictadura civil, resaltando el gobierno autoritario de Viktor Orbán y el deterioro del sistema de gobierno. Otro caso que podemos destacar es el de Venezuela, la cual sufre un retroceso democrático a partir de 2016; pasando de ser considerada una democracia presidencial a una dictadura civil. Sin embargo, hay casos como Liberia que sucede todo lo contrario, destacando el avance en el proceso democrático mencionado en la figura anterior y dejando de ser considerada una dictadura civil para ser considerada como una democracia presidencial.

Si observamos los cambios en el tipo de elecciones en la Figura 4.5, podemos ver que Venezuela llegó a presentar hasta 3 categorías distintas: pasó de elecciones democráticas a elecciones no democráticas multi-partido y luego a tener elecciones de un sólo partido, reflejando la catástrofe sucesiva del sistema electoral bajo el régimen de Maduro. Otro ejemplo que podemos destacar

es Turquía, que pasa de tener elecciones democráticas a no democráticas multi-partido, en línea con el creciente autoritarismo del presidente Erdoğan.

Por último, la Figura 4.6 nos muestra si algún país ha sufrido cambios en diferentes variables políticas. Además de los países ya mencionados anteriormente como Venezuela, Turquía o Nicaragua, los cuales han sufrido cambios en todas las variables políticas, cabe destacar el caso de Comoras, que también presenta múltiples cambios en estas variables. Estos cambios se deben a un período de inestabilidad política, en las que se aplicaron cambios en el sistema como la reforma constitucional de 2018 y las elecciones de 2019, que concentraron el poder y limitaron la oposición.

Estas figuras refuerzan la idea de que varios países han experimentado retrocesos democráticos significativos, especialmente en torno a elecciones libres, alternancia en el poder y la clasificación del sistema de gobierno. Además, demuestran que estos cambios no son simultáneos: mientras algunos países cambian en 2016, otros lo hacen en 2018 o 2029, lo cual permite contextualizar algunos cambios políticos con eventos históricos concretos en cada nación. Este tipo de análisis temporal no solo nos permite identificar tendencias políticas, sino que también podemos asociar estos cambios a la percepción de felicidad de la población y evaluar si existe alguna relación.

Dado que uno de los objetivos principales de este análisis es estudiar la evolución de la felicidad y sus factores influyentes a lo largo del tiempo, necesitamos trabajar con una base de datos que tenga cobertura completa para el periodo 2015–2024. En este sentido, hemos optado por utilizar como base principal la combinación de las bases de datos World Happiness y Freedom in the World, ya que ofrecen la continuidad temporal necesaria para poder aplicar técnicas de medidas repetidas y estudiar cómo varían las observaciones de un mismo país a lo largo de los años. Además, la base de datos Freedom in the World contiene ciertas variables políticas con algunas diferencias notables en algunos casos, por lo que combinarla con la base de datos World Happiness nos puede servir de gran ayuda a la hora de ambientar ciertos cambios en la felicidad.

Por su parte, la base de datos Democracy Data presenta una limitación temporal importante, al contener información solo hasta el año 2020; además de que muchas de sus variables muestran poca o nula variación a lo largo del tiempo. Por ello, en lugar de excluir completamente esta base, la integramos como complemento, utilizándola para construir una fotografía del contexto institucional y político de cada país. Esta fotografía nos resulta especialmente útil a la hora de interpretar fenómenos detectados en el análisis longitudinal, como caídas o aumentos abruptos en el nivel de felicidad. En esos casos, el tipo de régimen, la alternancia en el poder o la existencia de elecciones libres pueden aportar información relevante para entender estas variaciones.

Una vez hemos preparado la base de datos final, podemos estudiar la evolución de la variable objetivo `happiness_score` entre 2015 y 2024. Para ello, recurriremos a herramientas visuales como mapas o gráficos para poder detectar tanto patrones y tendencias generales como variaciones específicas en determinados territorios. Estas visualizaciones nos pueden dar

ciertos indicios sobre algunas variables explicativas que podemos tener en cuenta de cara al análisis longitudinal.

4.3 Evolución de la felicidad a lo largo del tiempo

Con el propósito de determinar ciertas tendencias geográficas en los niveles de felicidad, utilizaremos mapas coropléticos que nos permiten visualizar espacialmente la distribución del `happiness_score`. Este método ayuda a detectar agrupaciones regionales, contrastes entre países vecinos y posibles comportamientos atípicos. Representando la información directamente sobre el territorio, los mapas complementan el análisis longitudinal a través de la contextualización geográfica de la puntuación de la felicidad.

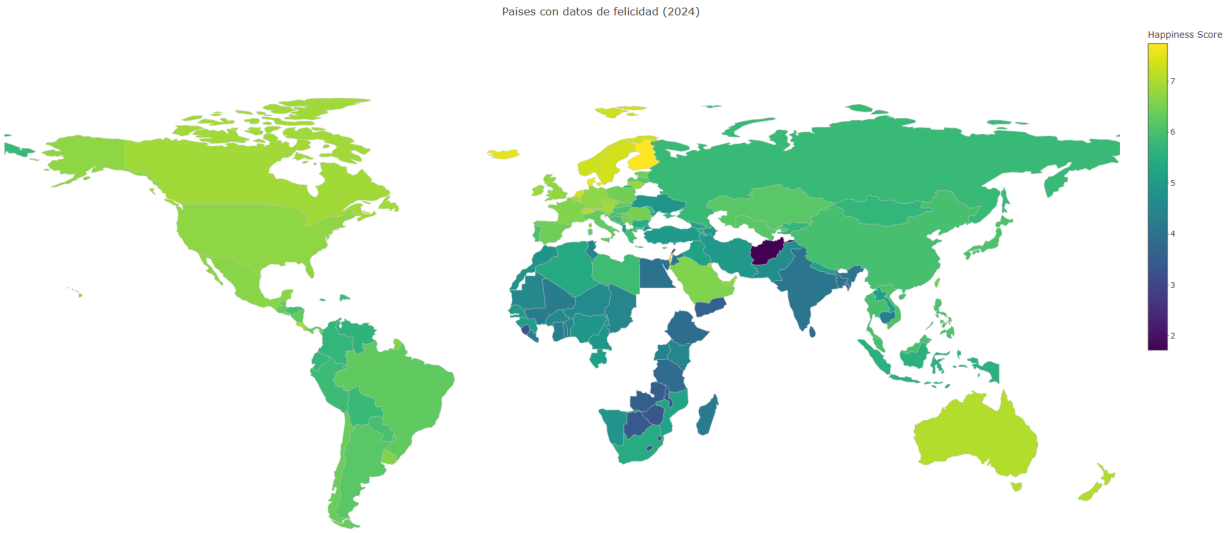

Figura 4.7: Representación de la puntuación global de felicidad en 2024.

El mapa de felicidad por país en 2024 de la Figura 4.7 muestra una clara diversidad geográfica en la percepción de la felicidad. Se observa que, por lo general, las regiones del hemisferio norte como Europa Occidental, América del Norte y Oceanía, tienen niveles más altos de felicidad (colores amarillos y verde claro), mientras que las regiones del hemisferio sur, especialmente África Subsahariana y algunas partes de Asia como Afganistán o India, presentan valores notablemente más bajos (colores azul oscuro o morado).

Países como Finlandia, Dinamarca e Islandia destacan con las puntuaciones más altas, demostrando estabilidad y altos niveles de desarrollo económico y social. En contraste, países como Zimbabwe, Líbano o Afganistán presentan los niveles más bajos, lo cual es coherente ya que son países donde rodeados por el conflicto, la pobreza y la inestabilidad política.

A través de este mapa no sólo diferenciamos diferencias entre regiones, sino que también podemos señalar patrones estructurales: Sudamérica, por ejemplo, muestra un nivel medio de felicidad, con cierta variabilidad entre países. Este mapa es útil cuando queremos detectar casos anómalos, como países con puntuaciones bajas en regiones generalmente altas o viceversa; como puede llegar a ser el caso de Afganistán, con una puntuación muy por debajo de sus países vecinos. En general, la figura muestra el impacto que pueden tener factores estructurales como el desarrollo económico o la estabilidad política en la percepción del bienestar de los ciudadanos a nivel global.

Además del análisis por país, resulta interesante analizar cómo cambia la felicidad entre regiones, y cómo se distribuye la felicidad en cada una de ellas. Para ello, utilizaremos diagramas de violín, que nos permiten observar la dispersión y concentración de los valores de `happiness_score`, además de su mediana y densidad; pudiendo analizar la distribución de la felicidad en cada región. Esta representación permite comparar el nivel de felicidad de las diferentes regiones y detectar posibles casos de desigualdades que podríamos no localizar en otro tipo de gráficos.

Observando la distribución de la felicidad por región de la Figura 4.8, si nos fijamos en la mediana (el rombo negro en cada violín), podemos contemplar cierta estabilidad en prácticamente todas las regiones; aunque hay regiones como Europa Occidental o Norteamérica y Australia que mantienen medianas altas a lo largo del tiempo, mientras que el Sur de Asia y África Sub-Sahariana tienen medianas más bajas.

En estos gráficos también se puede observar cierta bimodalidad, ya que algunas regiones, como Oriente Medio y África del Norte muestran dos modas (zonas más anchas del violín), lo que implicar cierta variabilidad interna: algunos países con altos niveles de felicidad conviven con otros con puntuaciones mucho más bajas. Por ejemplo, Asia del Sur presenta una distribución muy desigual en 2024, con un ensanchamiento en la parte inferior del gráfico que corresponde con una felicidad baja en ciertos países (como Afganistán).

En general, la forma de las distribuciones se mantiene similar entre 2015 y 2024 en muchas regiones, pero hay algunas en las que apreciamos mayor dispersión como es el caso Oriente Medio y Asia del Sur, los cuales sufren una caída considerable en la felicidad. Al contrario de estos últimos, hay otras regiones como Europa Occidental y Norteamérica y Australia que presentan distribuciones densas con una felicidad más alta y estable, lo que refleja mayor consistencia en el bienestar de sus ciudadanos.

A continuación analizamos la percepción de la corrupción, una variable que diversos estudios han señalado como influyente en la representación de la felicidad. En la Figura 4.9 se representa la distribución de esta variable mediante diagramas de violines, lo que nos permite observar también las diferencias entre regiones. Podemos analizar que hay algunas regiones

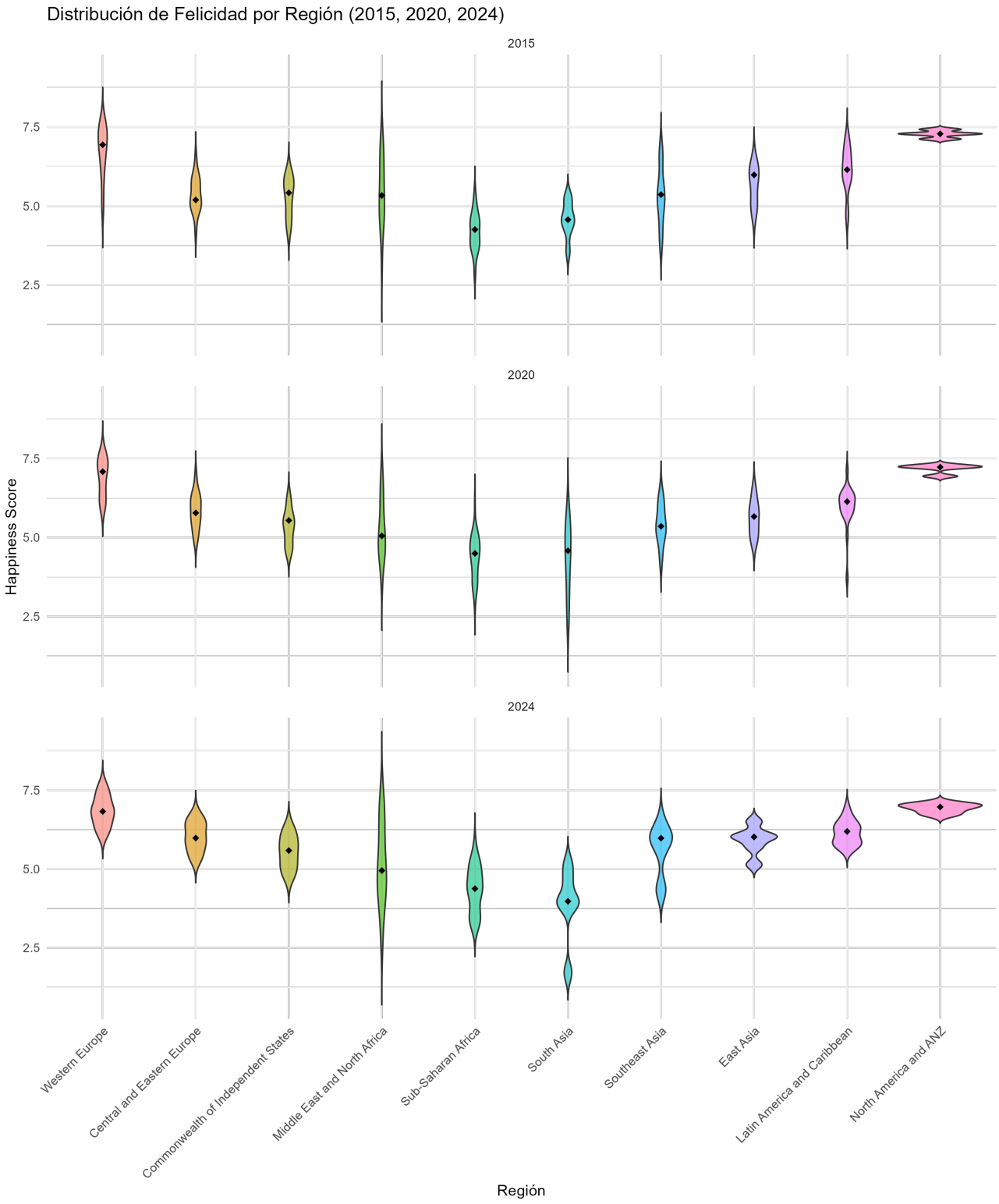

Figura 4.8: Distribución de la felicidad agrupada por región en los años 2015, 2020 y 2024.

que presentan niveles percibidos de corrupción notablemente más bajos, lo que podría estar asociado a mayores niveles de bienestar.

El diagrama de violines de la Figura 4.9 permite observar la evolución de la percepción de corrupción en las distintas regiones a lo largo del tiempo, destacando una clara bimodalidad en algunas regiones a partir de 2020 y especialmente en 2024, lo que sugiere una variabilidad creciente entre países dentro de la misma región.

Europa Occidental mantiene una distribución relativamente estable a lo largo del tiempo, destacando que siempre tiene países con una percepción de corrupción baja y otros con una percepción de la corrupción alta; con un valor de la mediana que sugiere una distribución simétrica. Sin embargo, en Europa Central y Oriental se observa un patrón inverso. En 2015 y 2020, la percepción es bastante alta, pero en 2024 se evidencia una bajada considerable en la percepción de la corrupción, lo que indica una mejoría de la percepción ciudadana sobre la corrupción en algunos países.

La Comunidad de Estados Independientes muestra una bimodalidad clara en 2024, mientras que en 2015 y 2020 tenía una distribución más concentrada. Esto se debe a que ahora hay países como Ucrania (0.04) que tienen niveles de percepción de corrupción muy bajos, mientras que el resto de países tienen una percepción de corrupción más alta. Sudamérica también muestra un cambio claro: de una percepción moderadamente alta en 2015 y 2020 a una percepción mucho más baja en 2024, aunque también podemos apreciar cierta bimodalidad debido a la baja corrupción percibida en países como Jamaica (0.05), mientras todo lo contrario ocurre en países como El Salvador (0.44).

Este análisis permite estudiar no solo la evolución de la percepción de la corrupción a lo largo del tiempo sino también la creciente variabilidad dentro de algunas regiones, lo que puede ser de utilidad para determinar si ciertos eventos políticos que desemboquen en una concentración del poder pueden estar afectando a una mayor percepción de la corrupción entre los ciudadanos.

Con el objetivo de comprender la evolución de la situación global en torno a diferentes variables políticas, analizamos a continuación la evolución promedio anual de las principales variables del informe de felicidad haciendo una comparación con 3 países que hemos destacado anteriormente por sus cambios políticos en los últimos años como son Hungría, Turquía y Venezuela. Esta visualización nos permitirá detectar posibles tendencias e impactos globales, como crisis políticas, sanitarias o económicas.

La evolución anual del promedio de las variables de interés de la Figura 4.10 refleja tendencias globales relevantes, y comparando su evolución con la de ciertos países concretos podemos determinar ciertas dinámicas y comprender cómo ciertos contextos políticos específicos pueden afectar a la felicidad de la población.

En primer lugar, Venezuela se desmarca de forma clara en varias variables. Mientras que a partir de 2017 el PIB per cápita se mantiene considerablemente constante, el de Venezuela empeiza a bajar, destacando una caída drástica en 2021, producto de la poca producción petrolera y la hiperinflación, reflejando la tremenda crisis económica que atraviesa el país.

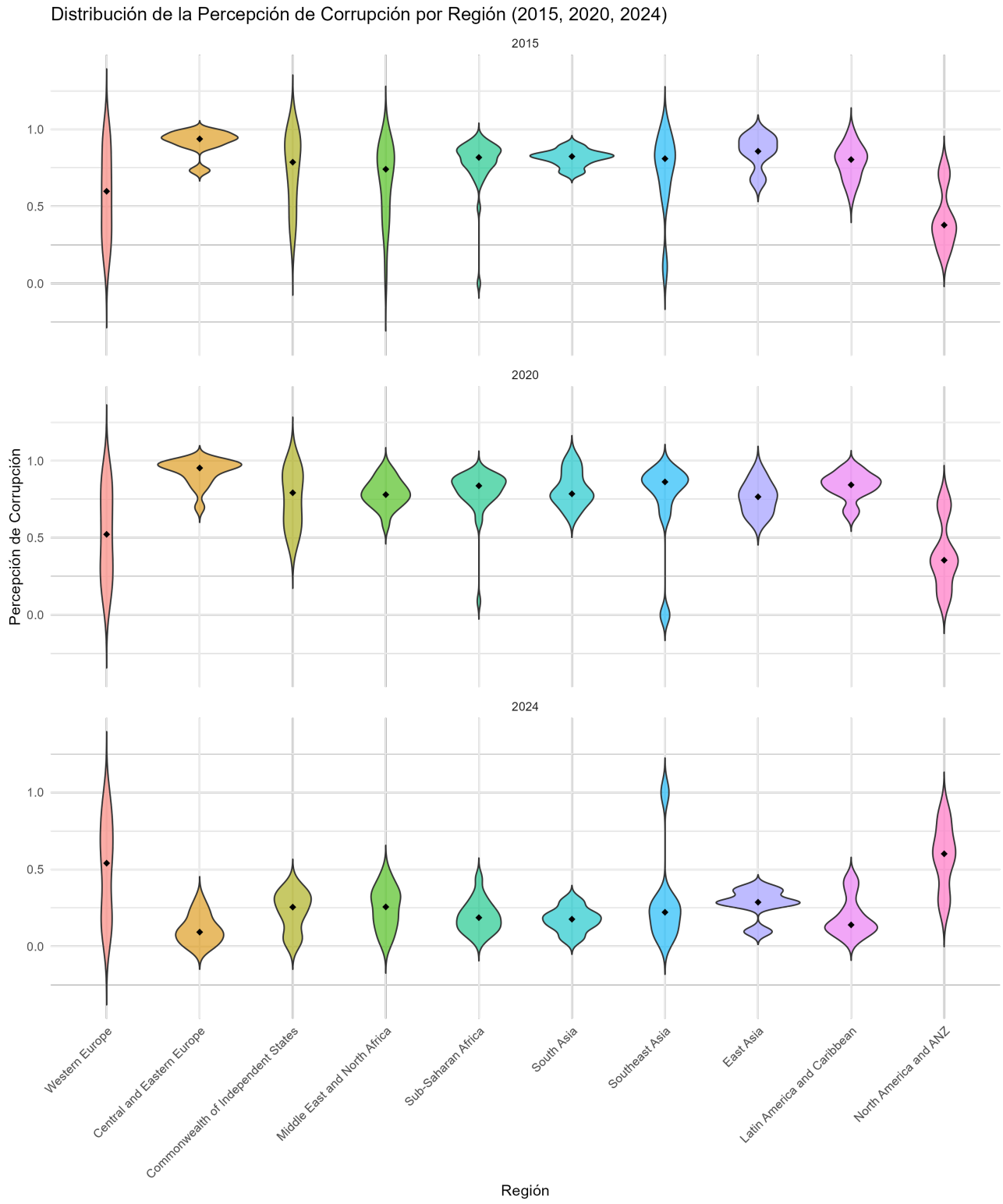

Figura 4.9: Distribución de la percepción de corrupción agrupada por región en los años 2015, 2020 y 2024.

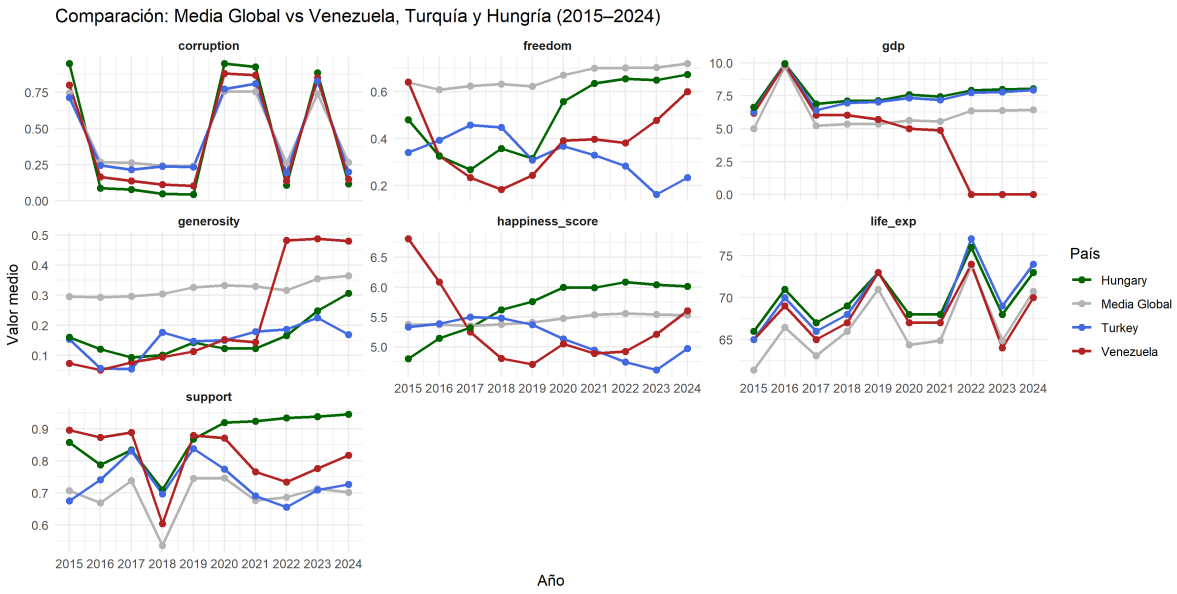

Figura 4.10: Análisis temporal de las principales variables del informe de felicidad en comparación con Hungría, Turquía y Venezuela.

Cabe destacar el descenso pronunciado que se produce en la libertad para tomar decisiones y en la felicidad percibida desde 2015, lo que podía sugerir un deterioro generalizado del bienestar, pero a partir de 2019 ambas variables empiezan a subir; efecto que puede haberse dado después de que las personas más descontentas formasen parte de la migración masiva que sufrió el país en años anteriores. Dicho aumento coincide con el mismo que sufre la generosidad a partir de 2019, superando incluso a la media global, lo cual podría estar relacionado a la solidaridad comunitaria surgida ante la crisis.

Por otro lado, aunque la libertad para tomar decisiones vaya creciendo cada año, Turquía muestra una tendencia claramente decreciente, especialmente desde 2018, coincidiendo con el fortalecimiento del poder ejecutivo de Recep Tayyip Erdoğan después de las elecciones. A pesar de mantener un PIB per cápita relativamente estable y una esperanza de vida en línea con la media global, la generosidad se mantiene bastante baja con respecto a la media, lo que, unido a la bajada de la felicidad a partir de 2019, puede reflejar un descontento ciudadano con la situación actual.

En el caso de Hungría, se observa una situación más confusa. Por un lado, variables como la esperanza de vida, el apoyo social y la felicidad percibida se mantienen por encima de la media global, lo que sugiere una cierta estabilidad a nivel civil. No obstante, la percepción de corrupción es elevada y la generosidad es bastante baja, lo cual coincide con el retroceso democrático que está sufriendo el país estos últimos años. Este contraste demuestra cómo un país puede mantener ciertos niveles de bienestar mientras sus organismos democráticos se deterioran.

Este análisis nos permite comprender el estudio previo de la evolución global. Hemos observado que algunas variables como la libertad para tomar decisiones, la generosidad o la propia felicidad han ido creciendo a lo largo del tiempo de forma estable, mientras que otras como la esperanza de vida han sufrido grandes fluctuaciones, algunas posiblemente como efectos secundarios post-pandemia. También destaca un pico atípico en el PIB per cápita en 2016, probablemente debido a valores extremos ocasionados por mala imputación en la recolecta de datos ya que hay muchos países en 2016 que presentan un PIB de 10, como es el caso de Ecuador: pasa de tener un PIB de 5.11 en 2015 a tener un PIB de 10 en 2016, para luego tener un PIB de 5.35 en 2017. También se observa una tendencia general de aumento en la felicidad percibida hasta 2022, seguida de un leve descenso. Por último, vemos que la percepción de corrupción también tiene ciertas oscilaciones entre años, lo que refleja fuertes diferencias entre países a lo largo del tiempo.

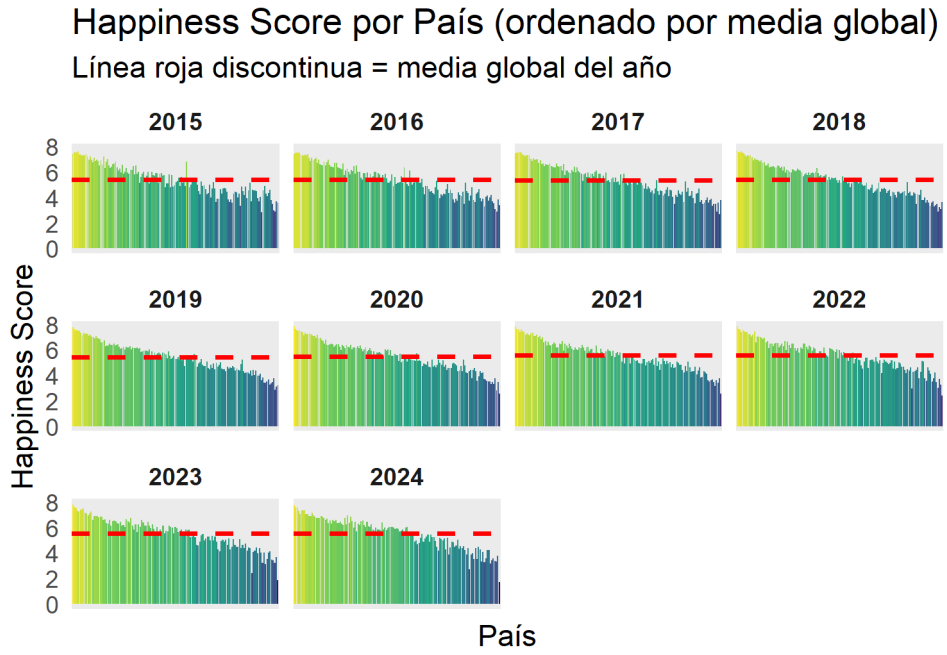

Figura 4.11: Análisis temporal de la evolución temporal de la felicidad por país.

La Figura 4.11 muestra la evolución del Happiness Score por país entre los años 2015 y 2024, en el que la línea roja discontinua indica la media global de felicidad de ese año. Además, para hacer un análisis coherente, en el que cada barra del gráfico representa el valor de felicidad para un país en un año determinado, haremos que las barras estén ordenadas de izquierda a derecha según la media de felicidad del país durante el periodo completo (2015–2024); es decir, los países más felices en promedio están situados a la izquierda en todos los años.

Podemos observar que la media global de felicidad se mantiene estable a lo largo de los años en un intervalo relativamente estrecho (entre 5.3 y 5.5 puntos), sugiriendo que la felicidad

mundial no ha sufrido ningún cambio contundente a lo largo de estos años. Sin embargo, esta estabilidad global en la puntuación no representa una estabilidad en todos los países.

A medida que progresan los años, especialmente a partir de 2020, observamos un aumento gradual en el número de países situados por debajo de la media global. Esta tendencia puede estar vinculada al impacto de la pandemia de COVID-19 y sus consecuencias, que afectaron, aunque no de la misma forma, a distintas regiones. En relación a lo visto anteriormente, esta tendencia también podría deberse a ciertos sucesos que han llevado al deterioro democrático y a la inestabilidad política en determinadas regiones.

La figura también demuestra las fuertes desigualdades en la distribución de la felicidad: mientras que un pequeño grupo de países (naciones del norte de Europa como Finlandia, Dinamarca o Islandia) se mantiene en la parte superior del ranking con puntuaciones muy por encima de la media, la mayoría de países se concentran en la parte media/baja. Esta distribución irregular indica que, aunque algunos países tienen niveles altos de felicidad y bienestar, muchos otros enfrentan obstáculos que les impiden mejorar su puntuación de felicidad; como Siria, Burundi o Afganistán.

Otro aspecto a tener en cuenta en la figura es el descenso en los valores máximos del Happiness Score en los últimos años: en 2023 y 2024, los países con los niveles más altos de felicidad han sufrido una pequeña bajada. Esta caída podría darse debido a acontecimientos globales que les han impactado más tarde, o simplemente como una normalización tras haber tenido niveles bastante altos los años anteriores.

En general, la Figura 4.11 no solo confirma que la media global se mantiene estable a lo largo de los años, sino que también muestra una dilatación de la brecha entre países que nos muestra la necesidad de buscar qué factores determinan esta diferencia de la felicidad entre países.

Tras analizar la evolución global de la felicidad, resulta interesante poner el foco en regiones más delimitadas que, por su cercanía, puedan mostrar dinámicas particulares a la hora de analizar la evolución de la felicidad en España. Con este fin, hemos seleccionado un grupo de países europeos próximos a España para realizar un análisis comparativo más detallado de la evolución del bienestar entre 2015 y 2024. Estos países son Portugal, Francia, Italia, Andorra, Suiza, Alemania y Reino Unido.

4.4 Evolución del Happiness Score en España

En este apartado queremos observar dos cosas: por un lado, estudiar la evolución de la felicidad en España a lo largo del tiempo; y por otro, comparar dicha evolución con la de sus países vecinos. Además, se analizarán variables como el PIB per cápita, el apoyo social y la libertad en la toma de decisiones para identificar posibles factores que describan esta evolución.

A través de esta perspectiva, no solo evaluamos si España sigue la misma evolución que sus países vecinos, sino que también permite detectar tendencias de bienestar y fortalezas en Europa Occidental.

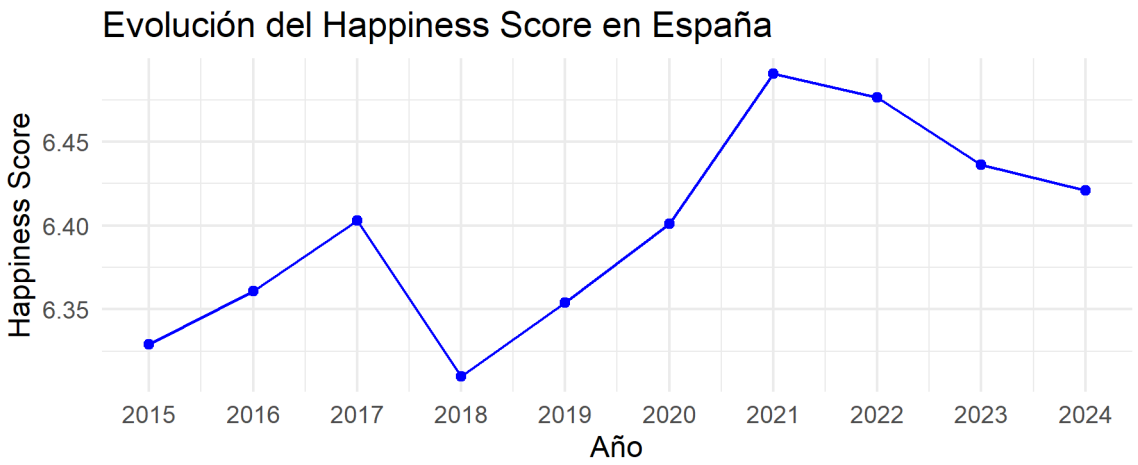

Figura 4.12: Evolución temporal de la felicidad en España.

La Figura 4.12, muestra una evolución relativamente estable de la puntuación de felicidad en España, con valores alrededor del 6.4. No obstante, destaca una bajada en 2018, seguida de una recuperación progresiva hasta 2021, el primer año después de la pandemia. Esta bajada puede deberse a los múltiples casos de corrupción que sacudieron el país y que terminaron con una moción de censura contra el presidente que supuso en un cambio de gobierno. Por el contrario, este máximo en 2021 podría deberse a la resiliencia que tuvo la ciudadanía ante la crisis sanitaria; o, viendo que en los años siguientes se produce un leve descenso de la felicidad, podría deberse a un efecto atrasado de las consecuencias de la pandemia.

La Figura 4.13 muestra la evolución de la felicidad en España en comparación con sus países vecinos. Lo primero que podemos observar es que España se sitúa en el rango medio-bajo del grupo, por encima únicamente de Portugal y, ocasionalmente, de Italia. Países como Suiza, Alemania y Reino Unido poseen holgadamente puntuaciones más altas (sobre todo Suiza), lo cual podría estar vinculado a mayores niveles de riqueza y estabilidad institucional.

Si analizamos la evolución de la felicidad, vemos que a partir de 2021 se produce un descenso claro en los países con niveles más altos de felicidad como Suiza o Alemania, mientras que Portugal sufre una gran subida y el resto de países se mantiene constante. Esto puede deberse a que ha habido una serie de factores, como ciertas tensiones geopolíticas, que han hecho afectado de manera directa a la felicidad de la población de muchos países.

Los gráficos presentados en las Figura 4.14, Figura 4.15 y Figura 4.16 permiten analizar la evolución de distintos factores que pueden haber afectado al bienestar y la calidad de vida de estos países.

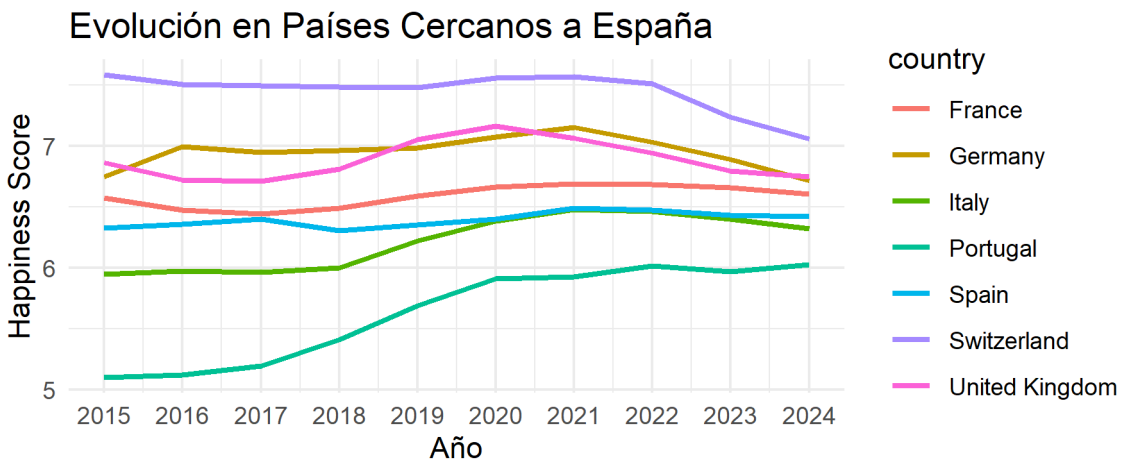

Figura 4.13: Evolución temporal de la felicidad en España y en países cercanos.

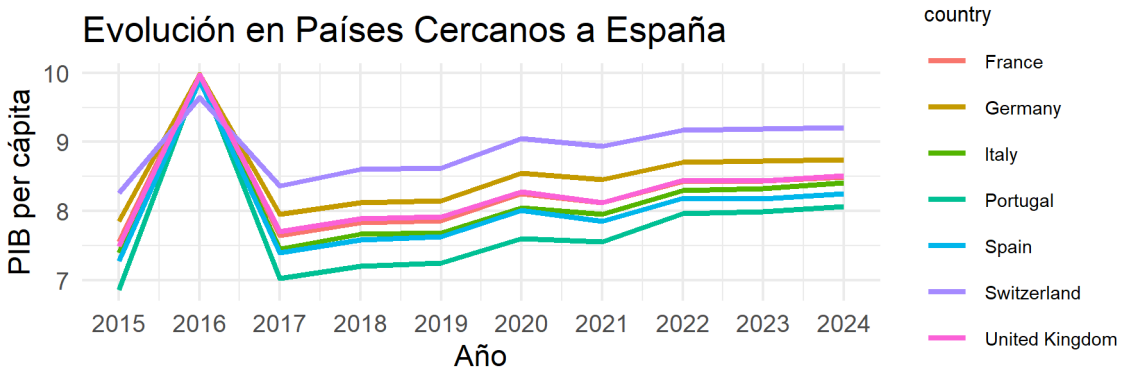

Figura 4.14: Evolución del PIB en España y sus países cercanos.

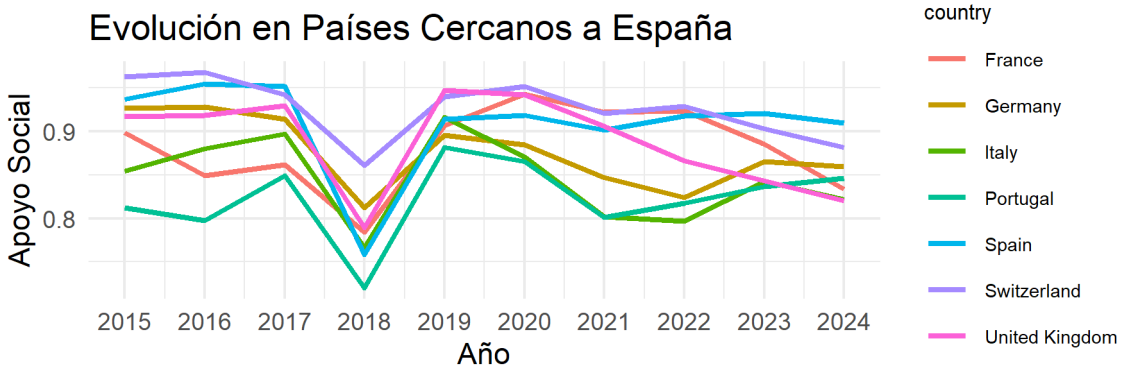

Figura 4.15: Evolución del apoyo social en España y sus países cercanos.

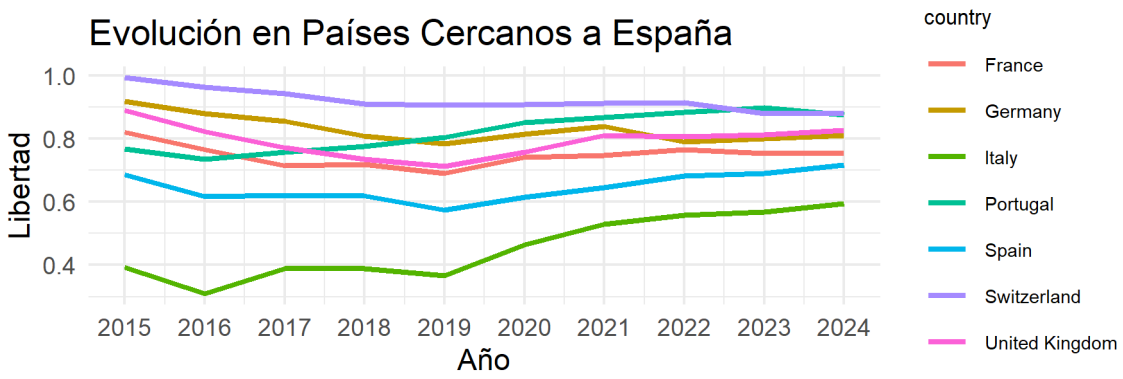

Figura 4.16: Evolución de la libertad en la toma de decisiones en España y sus países cercanos.

En la Figura 4.14, que muestra la evolución del PIB per cápita, se observa una fuerte caída en 2017, seguida por una recuperación progresiva desde 2018; la cual, como indiqué antes, puede deberse a una recodificación o mala recopilación de datos. En cualquier caso, destaca que Suiza mantiene sistemáticamente el mayor PIB per cápita, situándose de manera notoria por encima del resto del grupo. Le sigue Alemania, con valores elevados y estables, y un poco por debajo se encuentran Reino Unido y Francia. España se encuentra en la parte baja del grupo, sólo por encima de Portugal, aunque todos muestran una tendencia ascendente que refleja prosperidad económica.

En la Figura 4.15, relativo al apoyo social, se aprecia un comportamiento mucho menos constante. En 2018 hay una caída muy pronunciada en todos los países, especialmente en España, Italia y Portugal, lo cual puede deberse a una percepción negativa del entorno social como ocurrió en España debido a los casos de corrupción; o, al igual que con el PIB, a una recodificación de los datos. En 2019 se produce un repunte, y en los años siguientes se producen ciertas oscilaciones en la gran mayoría de países; destacando a España y Alemania como los países más constantes. No obstante, los cambios que se producen en el apoyo social no son comparables a los que se producen en la Figura 4.14 ya que no tienen la misma magnitud, y por lo general siempre tienen puntuaciones bastante altas.

La Figura 4.16 muestra la evolución de la libertad para tomar decisiones, la cual muestra valores mucho más dispares que la Figura 4.15. Aquí destaca el caso de Suiza, que mantiene los valores de libertad más altos durante todo el período, y el de Portugal, que sufre un gran ascenso que lo coloca en la parte más alta. España parte de un nivel más bajo en 2015 y experimenta una caída hasta 2019, seguida de una recuperación constante a partir de ese momento manteniéndose en la parte baja de la tabla. Italia, pese a ser el países con puntuación más baja, muestra una evolución ascendente sostenida, sobre todo a partir de 2019. Esta variable presenta una mejora generalizada en la región, manifestando ciertos avances en la percepción de libertad individual.

A través de estas figuras podemos sacar varias conclusiones. En primer lugar, España tiende a

situarse en posiciones relativamente bajas dentro del grupo de comparación, aunque muestra señales de mejora especialmente en PIB y libertad individual. Por otro lado, países como Suiza y Alemania destacan por sus buenos resultados en todos los ámbitos, lo cual explica por qué son los países con mayor puntuación de felicidad. Esta comparación demuestra que el análisis de medidas repetidas nos permite comprender mejor los factores que condicionan la evolución de la felicidad en países cercanos pero diferentes en muchos ámbitos.

En resumen, en este capítulo se ha realizado un análisis exploratorio exhaustivo de la base de datos utilizada, enfocándonos en la evolución temporal y variabilidad geográfica del índice de felicidad, así como su asociación a diferentes variables políticas, sociales y económicas. A través de herramientas visuales y comparaciones entre diferentes países, hemos reconocido ciertas tendencias relevantes en la representación de la felicidad. No obstante, todo lo hecho hasta ahora supone una fase meramente descriptiva del análisis. A partir del siguiente capítulo, daremos un paso más allá para intentar explicar el comportamiento del índice de felicidad, construyendo modelos estadísticos que nos ayuden a determinar qué variables influyen de forma significativa en su evolución y cómo interactúan entre sí.

5 Construcción del modelo

En este capítulo desarrollaremos el procedimiento de construcción de diferentes modelos estadísticos con el objetivo de explicar y predecir la percepción de felicidad (Happiness Score) a lo largo del tiempo, utilizando los datos longitudinales de la base de datos explicada en el capítulo anterior.

El propósito del capítulo es construir un modelo que sea estadísticamente sólido y, al mismo tiempo, interpretable desde el punto de vista de los diferentes factores sociales, económicos y políticos que puedan influir a la felicidad. Para ello, combinaremos técnicas de la estadística clásica (regresión lineal múltiple) con modelos diseñados específicamente para datos longitudinales, como los modelos lineales mixtos (LMM) y los modelos lineales generalizados mixtos (GLMM), que permiten capturar adecuadamente la estructura jerárquica de los datos y la dependencia entre medidas repetidas.

El modelado clásico se abordará desde dos estrategias complementarias:

- **Top-down**: partimos de un modelo completo que incluye todas las variables relevantes, y vamos eliminando aquellas que consideremos que no aportan información significativa o que generan problemas como multicolinealidad o sobreajuste.
- **Bottom-up**: comenzamos con un modelo simple con pocas variables y vamos añadiendo progresivamente nuevas variables explicativas, evaluando si su incorporación mejora el ajuste del modelo.

Ambas estrategias nos permiten explorar distintos métodos de construcción del modelo y encontrar un equilibrio entre simplicidad, robustez y capacidad explicativa.

En cambio, para los modelos mixtos (LMM y GLMM) seguimos un enfoque orientado a la validación: partimos directamente de estructuras con efectos aleatorios y efectos fijos determinados, evaluando su adecuación con herramientas gráficas y pruebas estadísticas. En el caso del LMM se emplea una estructura con todas las variables como efectos fijos y `year` como efecto aleatorio, mientras que para el GLMM se exploran múltiples combinaciones hasta identificar un modelo válido con distribución Gamma y `regional_indicator` como efecto aleatorio.

Dado que trabajamos con datos que varían a lo largo del tiempo para cada país, es importante distinguir entre:

- **Variables longitudinales**: cambian con el tiempo (por ejemplo, `gdp`, `support`, `freedom`, `generosity`, `life_exp`, `corruption`).
- **Variables fijas**: son aquellas que no cambian a lo largo del tiempo dentro del periodo de análisis, o que se consideran características estructurales del país. Como mencionamos anteriormente, estas variables se utilizan como contexto porque aportan información relevante sobre el entorno político o geográfico en el que se sitúa cada observación. Por ejemplo, `region` indica la ubicación geográfica del país y puede influir en aspectos culturales, económicos o sociales; al igual que las variables políticas que hemos incorporado como "foto" desde otras bases de datos como `is_democracy`, `regime_category`, `has_free_and_fair_election` o `has_alternation`.

En cuanto a la organización del capítulo, comenzaremos estudiando la relación entre las distintas variables de nuestra base de datos, revisando lo observado en el capítulo anterior pero con vistas a sacar posibles hipótesis sobre el posible efecto que tendrán en los modelos. Después, describiremos los criterios de selección de variables y la estrategia general de modelado en el caso de la regresión lineal múltiple; seguido del diagnóstico y validación del modelo clásico resultante. Posteriormente, se abordarán los modelos lineales mixtos (LMM) y los modelos lineales generalizados mixtos (GLMM), detallando aquellos que resultaron válidos y se utilizaron para predecir la evolución del Happiness Score. El capítulo concluye con una comparación de los modelos ajustados y una reflexión sobre sus implicaciones prácticas.

5.1 Análisis exploratorio y selección inicial de variables

Aunque ya hemos analizado previamente la estructura de los datos, antes de ajustar los modelos es útil revisar de nuevo las características más relevantes de cara a identificar posibles variables que podrían actuar como buenos predictores del `happiness_score`.

5.1.1 Estructura del dataset longitudinal

Nuestro conjunto de datos contiene observaciones anuales de múltiples países en el período 2015–2024, y está definido a través de variables socioeconómicas y de diferentes factores políticos. Las variables disponibles, descritas en el capítulo anterior, son: `regional_indicator`, `gdp`, `support`, `life_exp`, `freedom`, `generosity`, `corruption`, `status`, `political_rights`, `civil_liberties`, `fair_election`, `regime_category`, `democracy`, `electoral_category`, `presidential`, `alternation` y `year`.

Dado el planteamiento longitudinal de nuestro estudio, empleamos `country` como unidad de agrupación para modelar efectos aleatorios específicos de cada país a lo largo del tiempo, captando así variaciones propias de la evolución de cada país. Por otro lado, la variable `year` se incluye como efecto fijo porque el tiempo es un elemento compartido por todos los países; aunque su uso como efecto aleatorio se explora en los modelos mixtos más complejos ya que

modelando la variabilidad entre años podemos captar diferentes factores que afectan a la felicidad en cada uno de ellos. De la misma manera, la variable `regional_indicator` también se incluye como efecto aleatorio ya que modelando la variabilidad entre regiones podemos tener en cuenta las diferentes dependencias estructurales entre países que pertenecen a la misma región.

5.1.2 Visualización y evolución temporal de las variables

Antes de proceder al ajuste del modelo, vamos a recordar cómo evolucionan algunas de las principales variables explicativas a lo largo del tiempo. En la Figura 5.1 se muestran los gráficos de evolución de tres de las variables más relevantes: `happiness_score`, `corruption` y `generosity`. Hemos seleccionado estas dos últimas variables ya que consideramos que pueden tener cierta influencia en la evolución de la felicidad a lo largo del tiempo.

Esta visualización nos permite identificar posibles tendencias globales a lo largo del tiempo, lo cual resulta clave para elegir la estructura con la que construiremos nuestros modelos más adelante.

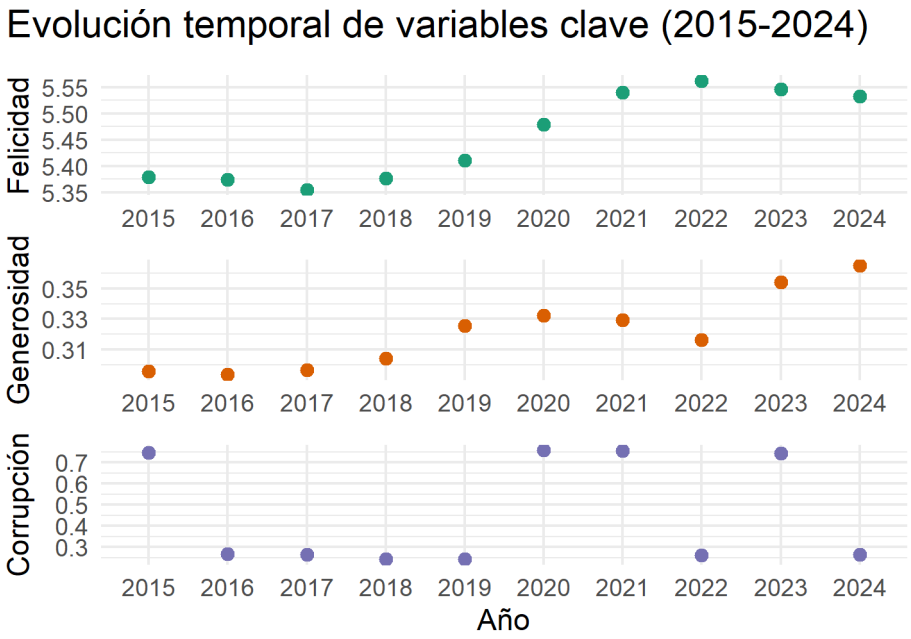

Figura 5.1: Evolución de la media anual de las variables felicidad, generosidad y percepción de corrupción (2015-2024).

Observando la Figura 5.1, vemos que la generosidad (`generosity`) aumenta moderadamente en los últimos años, mientras que la percepción de corrupción (`corruption`) sufre fluctuaciones

abruptas que no nos permiten detectar una tendencia clara. La puntuación de felicidad (`happiness_score`) se mantiene notablemente estable en el tiempo, y, si comparamos su evolución con la de la `generosity`, podemos ver que tienen una evolución bastante similar; por lo que la generosidad puede tener cierto impacto en la evolución de la felicidad. Viendo la evolución de la corrupción, vemos que es necesario utilizar modelos que capten la variabilidad de cada país, ya que el promedio global puede estar ocultando comportamientos diferentes en algunos países.

5.1.3 Matriz de correlaciones

Antes de ajustar ningún modelo, es útil examinar la correlación entre las variables explicativas numéricas y la variable objetivo `happiness_score`. Esto nos permitirá identificar posibles relaciones lineales, evaluar redundancias y tomar decisiones informadas sobre qué variables incluir inicialmente en el modelo.

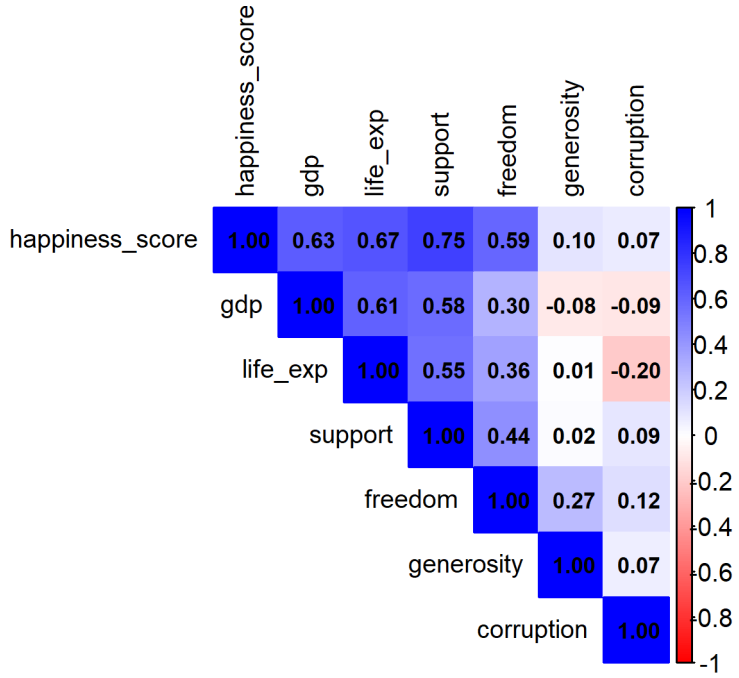

Figura 5.2: Matriz de correlaciones de las variables numéricas de nuestra base de datos.

Como observamos en la Figura 5.2, las correlaciones más fuertes con `happiness_score` provienen de:

- `support` (0.75): el apoyo social es la variable que más se relaciona con la felicidad.
- `life_exp` (0.67) y gdp (0.63): muestran correlaciones altas, mostrando la importancia del bienestar económico y la salud en la evolución de la felicidad.

- `freedom` (0.59) también presenta una correlación moderada.

En cambio, variables como `generosity` (0.10) y `corruption` (0.07) muestran una relación muy débil con la felicidad, lo que nos hace cuestionar su importancia explicativa. Las correlaciones entre variables explicativas no son excesivamente altas (ninguna supera 0.8), por lo que, en principio, no deberíamos de tener muchos problemas de multicolinealidad. Aun así, esto se verificará con el cálculo del Variance Inflation Factor (VIF), un indicador que mide cuánto aumenta la varianza de los coeficientes estimados por la multicolinealidad. Un valor de VIF superior a 5 suele considerarse problemático, ya que sugiere una multicolinealidad como mínimo moderada, lo que puede afectar a la estabilidad e interpretación de los coeficientes. En este trabajo utilizamos el VIF para asegurarnos que las variables seleccionadas no tienen demasiada correlación entre ellas.

Si nos guiáramos exclusivamente por la matriz de correlaciones, deberíamos de incluir `support`, `life_exp`, `gdp` y `freedom` en el modelo, pero estas correlaciones deben interpretarse con cautela ya que se calculan sobre medidas repetidas para los mismos países a lo largo del tiempo; lo que puede llevar a una fuerte correlación debido a la estructura longitudinal de los datos.

5.2 Criterios de selección del modelo

Al construir un modelo estadístico, nos encontramos con el caso de que contamos con múltiples combinaciones posibles de variables explicativas. Para elegir la combinación más adecuada, se utilizan criterios que balancean dos aspectos fundamentales: el ajuste al conjunto de datos (lo bien que predice el modelo los datos observados), y la complejidad del modelo (cuántos parámetros se incluyen). Dos de los criterios más utilizados para este propósito son el Akaike Information Criterion (AIC) y el Bayesian Information Criterion (BIC).

AIC penaliza la complejidad del modelo y busca minimizar la pérdida de información. Se calcula como:

$$\text{AIC} = -2 \cdot \log(\hat{L}) + 2k.$$

donde $\hat{L}$ es la verosimilitud máxima del modelo y k el número de parámetros.

BIC penaliza de manera más severa los modelos complejos (dependiendo del tamaño de muestra n):

$$\text{BIC} = -2 \cdot \log(\hat{L}) + k \cdot \log(n).$$

Un AIC o BIC más bajo indica un mejor modelo, pero el AIC tiende a ser más flexible con modelos más complejos; mientras que el BIC favorece modelos más sencillos.

5.3 Modelado clásico

Como punto de partida, partiremos de un modelo clásico de regresión lineal múltiple para explicar el `happiness_score` a partir de variables como `gdp`, `life_exp`, `support`, `freedom`, `generosity` y `corruption`. Para explorar la mejor combinación de predictores, aplicamos dos estrategias de selección de variables:

5.3.1 Estrategia top-down (backward elimination)

Usamos el criterio AIC y BIC para eliminar aquellas variables cuya exclusión mejora la simplicidad del modelo sin sacrificar capacidad predictiva. A continuación, se muestra la salida completa generada por R utilizando los criterios AIC y BIC, respectivamente:

```
Start:  AIC=-1590.33
happiness_score ~ gdp + life_exp + support + freedom + generosity +
    corruption

             Df Sum of Sq    RSS     AIC
<none>                    496.37 -1590.3
- generosity  1     2.207 498.57 -1585.8
- corruption  1    10.559 506.93 -1561.3
- gdp         1    35.828 532.19 -1489.6
- life_exp    1    73.315 569.68 -1389.3
- freedom     1    88.735 585.10 -1349.9
- support     1   132.429 628.80 -1243.7

Start:  AIC=-1553.26
happiness_score ~ gdp + life_exp + support + freedom + generosity +
    corruption

             Df Sum of Sq    RSS     AIC
- generosity  1     2.207 498.57 -1554.0
<none>                    496.37 -1553.3
- corruption  1    10.559 506.93 -1529.5
- gdp         1    35.828 532.19 -1457.8
- life_exp    1    73.315 569.68 -1357.5
- freedom     1    88.735 585.10 -1318.1
- support     1   132.429 628.80 -1212.0

Step:  AIC=-1554.02
happiness_score ~ gdp + life_exp + support + freedom + corruption
```

```
             Df Sum of Sq    RSS     AIC
<none>                     498.57 -1554.0
- corruption  1    10.754 509.33 -1529.9
- gdp         1    34.143 532.72 -1463.7
- life_exp    1    73.635 572.21 -1358.3
- freedom     1   106.041 604.61 -1277.1
- support     1   131.548 630.12 -1216.2
```

En la selección según AIC, el modelo presenta un AIC final de -1590.3, donde no se excluye ninguna variable. Por su parte, en la selección según BIC, en la que se penalizan más los modelos complejos, el modelo óptimo presenta un BIC de -1554, descartando `generosity`. El hecho de que ambos modelos no coincidan muestra que el modelo no es lo suficientemente robusto frente a cambios en la penalización por complejidad; lo cual nos puede llevar a problemas con nuestros datos longitudinales, y consolida la necesidad de emplear métodos más adecuados para esta estructura como los modelos mixtos.

5.3.2 Estrategia bottom-up (forward selection)

También incorporamos una estrategia bottom-up (selección hacia adelante), que parte de un modelo nulo y añade variables una a una en función de la mejora del AIC. Esta estrategia permite comprobar si existe alguna combinación alternativa de predictores que produzca un modelo competitivo o incluso mejor al obtenido por eliminación hacia atrás.

```
Start:  AIC=369.95
happiness_score ~ 1

             Df Sum of Sq     RSS     AIC
+ support     1   1052.77  839.17 -826.33
+ life_exp    1    846.37 1045.57 -502.19
+ gdp         1    762.22 1129.72 -388.10
+ freedom     1    661.35 1230.60 -262.03
+ generosity  1     20.45 1871.50  355.93
+ corruption  1     10.22 1881.72  363.96
<none>                   1891.95  369.95

Step:  AIC=-826.33
happiness_score ~ support

             Df Sum of Sq    RSS      AIC
+ life_exp    1   179.860 659.31 -1179.89
```

```
+ freedom     1   165.587 673.59 -1148.32
+ gdp         1   116.925 722.25 -1045.50
+ generosity  1    15.313 823.86  -851.48
<none>                    839.17  -826.33
+ corruption  1     0.023 839.15  -824.37

Step:  AIC=-1179.89
happiness_score ~ support + life_exp

             Df Sum of Sq    RSS     AIC
+ freedom     1   117.915 541.40 -1468.3
+ gdp         1    31.208 628.11 -1249.4
+ corruption  1    20.373 638.94 -1224.2
+ generosity  1    15.743 643.57 -1213.5
<none>                    659.31 -1179.9

Step:  AIC=-1468.33
happiness_score ~ support + life_exp + freedom

             Df Sum of Sq    RSS     AIC
+ gdp         1    32.072 509.33 -1556.3
+ corruption  1     8.683 532.72 -1490.2
<none>                    541.40 -1468.3
+ generosity  1     0.642 540.76 -1468.1

Step:  AIC=-1556.34
happiness_score ~ support + life_exp + freedom + gdp

             Df Sum of Sq    RSS     AIC
+ corruption  1   10.7540 498.57 -1585.8
+ generosity  1    2.4018 506.93 -1561.3
<none>                    509.33 -1556.3

Step:  AIC=-1585.79
happiness_score ~ support + life_exp + freedom + gdp + corruption

             Df Sum of Sq    RSS     AIC
+ generosity  1    2.2067 496.37 -1590.3
<none>                    498.57 -1585.8

Step:  AIC=-1590.33
happiness_score ~ support + life_exp + freedom + gdp + corruption +
    generosity
```

En este caso, se fueron incorporando variables hasta alcanzar el modelo completo, es decir, con todas las variables explicativas. Este resultado coincide con el modelo obtenido por backward elimination por AIC, pero difiere en el seleccionado por BIC ya que este descartaba `generosity`. Esta diferencia muestra que los criterios de selección pueden llevar a soluciones distintas según el punto de partida.

5.3.3 Diagnóstico y validación final del modelo

En base a los resultados anteriores, el modelo final que utilizaremos para el diagnóstico y validación es el siguiente:

$$happiness\_score_{ij} = \beta_0 + \beta_1\, gdp_{ij} + \beta_2\, life\_exp_{ij} + \beta_3\, support_{ij} + \beta_4\, freedom_{ij} + \beta_5\, corruption_{ij} + \epsilon_{ij}.$$

```
Start:  AIC=369.95
happiness_score ~ 1

             Df Sum of Sq     RSS     AIC
+ support     1   1052.77  839.17 -826.33
+ life_exp    1    846.37 1045.57 -502.19
+ gdp         1    762.22 1129.72 -388.10
+ freedom     1    661.35 1230.60 -262.03
+ generosity  1     20.45 1871.50  355.93
+ corruption  1     10.22 1881.72  363.96
<none>                    1891.95  369.95

Step:  AIC=-826.33
happiness_score ~ support

             Df Sum of Sq    RSS      AIC
+ life_exp    1   179.860 659.31 -1179.89
+ freedom     1   165.587 673.59 -1148.32
+ gdp         1   116.925 722.25 -1045.50
+ generosity  1    15.313 823.86  -851.48
<none>                    839.17  -826.33
+ corruption  1     0.023 839.15  -824.37

Step:  AIC=-1179.89
happiness_score ~ support + life_exp

             Df Sum of Sq    RSS     AIC
+ freedom     1   117.915 541.40 -1468.3
+ gdp         1    31.208 628.11 -1249.4
```

```
+ corruption  1    20.373 638.94 -1224.2
+ generosity  1    15.743 643.57 -1213.5
<none>                   659.31 -1179.9

Step:  AIC=-1468.33
happiness_score ~ support + life_exp + freedom

             Df Sum of Sq    RSS     AIC
+ gdp         1    32.072 509.33 -1556.3
+ corruption  1     8.683 532.72 -1490.2
<none>                   541.40 -1468.3
+ generosity  1     0.642 540.76 -1468.1

Step:  AIC=-1556.34
happiness_score ~ support + life_exp + freedom + gdp

             Df Sum of Sq    RSS     AIC
+ corruption  1   10.7540 498.57 -1585.8
+ generosity  1    2.4018 506.93 -1561.3
<none>                   509.33 -1556.3

Step:  AIC=-1585.79
happiness_score ~ support + life_exp + freedom + gdp + corruption

             Df Sum of Sq    RSS     AIC
+ generosity  1    2.2067 496.37 -1590.3
<none>                   498.57 -1585.8

Step:  AIC=-1590.33
happiness_score ~ support + life_exp + freedom + gdp + corruption +
    generosity
```

Para comprobar que no tenemos problemas de multicolinealidad, calcularemos el VIF del modelo. Podemos confirmar que, en efecto, ninguna de nuestras variables genera grandes problemas de multicolinealidad (**1.853, 1.957, 1.893, 1.295, 1.144**), ya que no contamos con ningún valor de VIF superior a 2. Como veremos en la siguiente sección, ahora procederemos a realizar el diagnóstico y validación final del modelo, verificando si cumple con las hipótesis necesarias para garantizar dicha validez. Estas hipótesis incluyen:

- Normalidad de los residuos.
- Media cero de los residuos.
- Homoscedasticidad (varianza constante de los errores).

- Independencia de los errores.

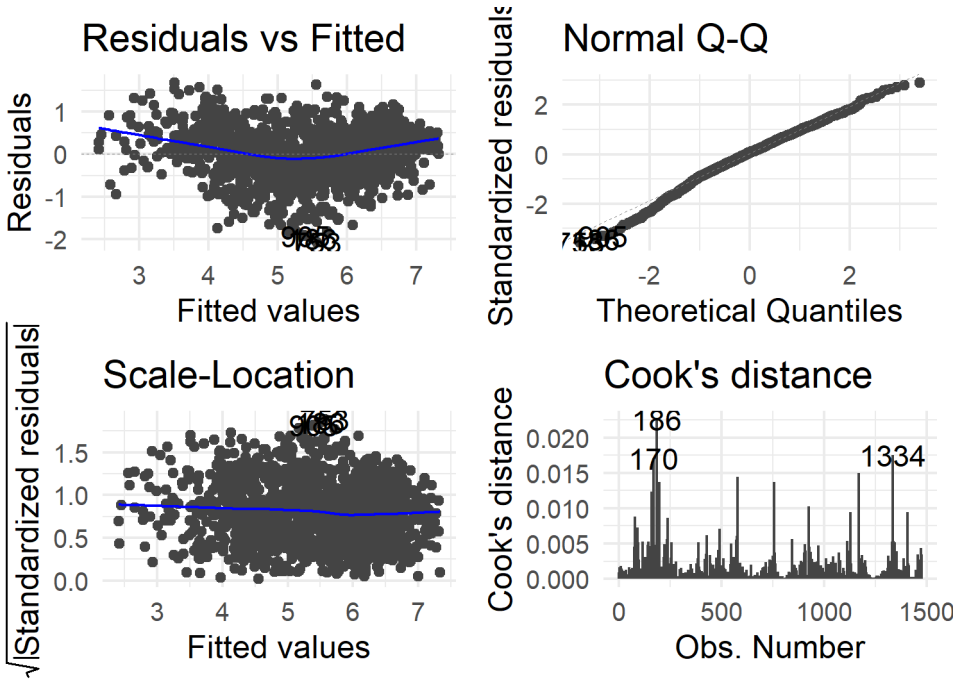

Figura 5.3: Gráficas para la validación del modelo clásico.

5.3.3.1 Normalidad de los residuos

Si observamos el gráfico Q-Q de la Figura 5.3, llega un momento en el que los residuos se desvían de la línea teórica. No obstante, aplicamos el test de Jarque-Bera para evaluar de forma numérica si los residuos del modelo siguen una distribución normal. El resultado indica un p-valor muy bajo ($<$**0.001**), lo que nos lleva a rechazar la hipótesis de normalidad.

5.3.3.2 Media cero

De la misma manera, podemos calcular de forma numérica la media de los residuos para verificar si se aproxima a cero, como requiere el modelo. En este caso, se cumple adecuadamente (**0**). Sin embargo, esta no es la forma correcta de determinar la media cero de los residuos, sino que hay que hacer una interpretación gráfica. Si observamos la gráfica Residuals vs Fitted en la Figura 5.3, podríamos decir que los valores más bajos se alejan del eje de abscisas, por lo que se cumple los residuos no tienen media cero.

5.3.3.3 Homoscedasticidad

El gráfico de residuos frente a los valores ajustados de la Figura 5.3 muestra cierta dispersión irregular, por lo que podría haber heterocedasticidad. Para comprobarlo de forma numérica, se aplica la prueba de Breusch-Pagan para comprobar si la varianza de los errores es constante. El resultado del test devuelve un p-valor muy bajo (**<0.001**), lo que sugiere que, efectivamente, no hay varianza constante.

5.3.3.4 No correlación de los errores

Si observamos la gráfica Residuals vs Fitted en la Figura 5.3, podemos observar una cierta curva que nos pueda hacer pensar que los errores están correlacionados. Mediante el test de Durbin-Watson, verificamos de forma numérica que no exista esta autocorrelación en los residuos. En este caso, el p-valor vuelve a ser muy bajo (**<0.001**), lo que nos hace rechazar la hipótesis nula y, por tanto, ver que existe autocorrelación de los errores.

5.3.3.5 Conclusión del diagnóstico

Estos incumplimientos de los supuestos teóricos sugieren que el modelo no es válido para trabajar con datos longitudinales. La inadecuación del modelo clásico justifica el uso de modelos más robustos capaces de incorporar la estructura jerárquica de los datos, como los modelos lineales mixtos (LMM), los cuales permiten modelar efectos aleatorios por país, capturar la correlación entre medidas repetidas y mejorar la validez e interpretación de los resultados.

5.4 Modelos Lineales Mixtos (LMM)

En el modelado clásico mediante regresión lineal múltiple asumimos que las observaciones eran independientes entre sí, pero en nuestro caso trabajamos con datos longitudinales, es decir, con medidas repetidas a lo largo del tiempo. Esto conlleva a una dependencia entre observaciones que los modelos clásicos no pueden capturar adecuadamente, y para afrontar esta limitación recurrimos a los modelos lineales mixtos (LMM).

Los LMM permiten combinar efectos fijos, que capturan el efecto promedio de las variables explicativas sobre la variable respuesta, y efectos aleatorios, que permiten modelar la variabilidad específica entre países. De esta forma podemos capturar la estructura jerárquica de los datos, considerando que cada país puede tener un nivel base de felicidad distinto (intercepto propio), mientras que los efectos de las variables predictoras son comunes a todos los países. En nuestro modelo, consideramos efectos fijos aquellas variables explicativas cuyo efecto queremos estimar de forma general para toda la población (en este caso, todos los países y años). Estas variables incluyen `gdp`, `support`, `freedom`, `life_exp`, `corruption` y

las variables políticas (`is_democracy`, `regime_category`, etc.). También introducimos un intercepto aleatorio por país porque cada país tiene un nivel base distinto de felicidad no explicado por las variables fijas, hay dependencia entre observaciones del mismo país en distintos años, y no nos interesa estimar el efecto específico de cada país, sino tener en cuenta la variabilidad entre ellos. Además, se incorpora también `year` como efecto aleatorio, ya que se asume que la evolución temporal del `Happiness Score` no es idéntica en todos los países, ya que algunos pueden experimentar mejoras sostenidas a lo largo del tiempo, mientras que otros pueden estancarse o incluso empeorar. Permitir que la relación con el tiempo varíe entre países nos ayuda a modelar mejor esta variabilidad en las evoluciones temporales, respetando la estructura longitudinal de los datos. El modelo lineal mixto que vamos a plantear incluye todas las variables de nuestra base de datos como efectos fijos ya que vamos a asumir que todas tienen influencia en la felicidad. Nuestro LMM es el siguiente:

$$\begin{aligned} happinessscore_{ij} = \beta_0 &+ \beta_1 \cdot regionalindicator_{ij} + \beta_2 \cdot \text{gdp}_{ij} + \beta_3 \cdot \text{support}_{ij} + \beta_4 \cdot lifeexp_{ij} \\ &+ \beta_5 \cdot \text{freedom}_{ij} + \beta_6 \cdot \text{generosity}_{ij} + \beta_7 \cdot \text{corruption}_{ij} + \beta_8 \cdot \text{status}_{ij} \\ &+ \beta_9 \cdot politicalrights_{ij} + \beta_{10} \cdot civilliberties_{ij} + \beta_{11} \cdot fairelection_{ij} \\ &+ \beta_{12} \cdot regimecategory_{ij} + \beta_{13} \cdot \text{democracy}_{ij} + \beta_{14} \cdot electoralcategory_{ij} \\ &+ \beta_{15} \cdot \text{presidential}_{ij} + \beta_{16} \cdot \text{alternation}_{ij} + \beta_{17} \cdot \text{year}_{ij} \\ &+ u_{0i} + u_{1i} \cdot \text{year}_{ij} + \varepsilon_{ij}. \end{aligned}$$

donde:

- i representa el país y j el año.
- β_k son los coeficientes fijos asociados a cada variable explicativa.
- $u_{0i} \sim \mathcal{N}(0, \sigma^2_{u0})$ es el intercepto aleatorio por país.
- $u_{1i} \sim \mathcal{N}(0, \sigma^2_{u1})$ es el efecto aleatorio asociado al año dentro de cada país.
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ es el término de error residual.

Para seleccionar el mejor modelo, partimos del modelo mixto completo y generamos múltiples modelos candidatos con diferentes combinaciones de variables predictoras. Iniciamos una búsqueda automática que identifica los modelos con mejor equilibrio entre ajuste y complejidad, evaluado mediante el AIC. Recordemos que para comparar modelos con distintos efectos fijos, necesitamos ajustar el modelo inicial utilizando máxima verosimilitud (ML) en lugar de REML, ya que REML se utiliza cuando la estructura de efectos fijos es la misma. De esta manera, generamos una lista de modelos candidatos ordenados por su AIC, de manera que si el primer modelo de la lista no resulta válido en nuestro futuro diagnóstico, pasamos al siguiente modelo de la lista; y así suvesivamente hasta que encontremos un LMM válido. A continuación, se muestra la salida del LMM:

```
happiness_score ~ civil_liberties + electoral_category + freedom +
    gdp + life_exp + political_rights + regime_category + regional_indicator +
    status + support + year + (1 + year | country)
```

	Estimate	Std. Error
(Intercept)	0.324164364	7.552933852
civil_liberties	-0.085706718	0.028206322
electoral_categoryno elections	0.316988344	0.569139562
electoral_categorynon-democratic multi-party elections	-0.101173814	0.517014422
electoral_categorysingle-party elections	-0.129317136	0.577749162
freedom	1.105677827	0.096976533
gdp	0.004129873	0.005079433
life_exp	0.006873066	0.002001046
political_rights	-0.005281865	0.023568327
regime_categoryMilitary dictatorship	-0.484888243	0.181621420
regime_categoryMixed democratic	0.058916397	0.512188688
regime_categoryParliamentary democracy	0.491964838	0.512982203
regime_categoryPresidential democracy	0.108316702	0.519870834
regime_categoryRoyal dictatorship	0.439323470	0.231154641
regional_indicatorCommonwealth of Independent States	0.025284819	0.139357843
regional_indicatorEast Asia	-0.197676926	0.176556838
regional_indicatorLatin America and Caribbean	0.240652851	0.164965613
regional_indicatorMiddle East and North Africa	-0.104757067	0.133162217
regional_indicatorNorth America and ANZ	0.688070128	0.273939730
regional_indicatorSouth Asia	-0.950068059	0.233030796
regional_indicatorSoutheast Asia	-0.311388279	0.176401551
regional_indicatorSub-Saharan Africa	-0.643376248	0.122628095
regional_indicatorWestern Europe	0.028827313	0.073907493
statusNF	0.150746827	0.101146917
statusPF	0.101287923	0.063310348
support	0.685181773	0.077924287
year	0.001860154	0.003774405

	t value
(Intercept)	0.0429190
civil_liberties	-3.0385642
electoral_categoryno elections	0.5569607
electoral_categorynon-democratic multi-party elections	-0.1956886
electoral_categorysingle-party elections	-0.2238292
freedom	11.4014988
gdp	0.8130579
life_exp	3.4347361
political_rights	-0.2241086

```
regime_categoryMilitary dictatorship                    -2.6697745
regime_categoryMixed democratic                          0.1150287
regime_categoryParliamentary democracy                   0.9590291
regime_categoryPresidential democracy                    0.2083531
regime_categoryRoyal dictatorship                        1.9005609
regional_indicatorCommonwealth of Independent States     0.1814381
regional_indicatorEast Asia                             -1.1196220
regional_indicatorLatin America and Caribbean            1.4588062
regional_indicatorMiddle East and North Africa          -0.7866876
regional_indicatorNorth America and ANZ                  2.5117573
regional_indicatorSouth Asia                            -4.0770064
regional_indicatorSoutheast Asia                        -1.7652242
regional_indicatorSub-Saharan Africa                    -5.2465648
regional_indicatorWestern Europe                         0.3900459
statusNF                                                 1.4903749
statusPF                                                 1.5998636
support                                                  8.7929168
year                                                     0.4928338
```

La salida del modelo mixto lineal seleccionado muestra un ajuste con un gran número de efectos fijos y un efecto aleatorio de año por país, lo que permite capturar variaciones estructurales tanto entre países como a lo largo del tiempo. En concreto, el modelo es el siguiente:

$$\begin{aligned} happinessscore_{ij} = \beta_0 &+ \beta_1 \cdot civilliberties_{ij} + \beta_2 \cdot electoralcategory_{ij} + \beta_3 \cdot \text{freedom}_{ij} \\ &+ \beta_4 \cdot \text{gdp}_{ij} + \beta_5 \cdot lifeexp_{ij} + \beta_6 \cdot politicalrights_{ij} \\ &+ \beta_7 \cdot regimecategory_{ij} + \beta_8 \cdot regionalindicator_{ij} + \beta_9 \cdot \text{status}_{ij} \\ &+ \beta_{10} \cdot \text{support}_{ij} + \beta_{11} \cdot \text{year}_{ij} + u_{0j} + u_{1j} \cdot \text{year}_{ij} + \epsilon_{ij}. \end{aligned}$$

Entre los efectos fijos, se observa que `support`, `life_exp` y especialmente `freedom` tienen efectos positivos y estadísticamente significativos sobre el nivel de felicidad, lo que indica que mayores niveles de libertad para tomar decisiones, esperanza de vida y apoyo social están fuertemente asociados con una mayor puntuación de felicidad. También destaca el indicador regional North America and ANZ, con un coeficiente positivo y significativo, lo que indica que esta región tiene niveles de felicidad altos, mientras que otras regiones como South Asia o Sub-Saharan Africa presentan efectos negativos y significativos, evidenciando una menor felicidad media en esas áreas. Dentro de las variables políticas, el régimen de Military dictatorship tiene un efecto negativo relevante, mientras que el statusPF (partly free) tiene un efecto positivo aunque de menor magnitud. En definitiva, el modelo explica adecuadamente la variabilidad de la felicidad teniendo en cuenta factores estructurales, económicos, sociales y políticos.

Para evaluar la calidad del modelo, utilizaremos las siguientes medidas: el R^2 marginal, que representa la proporción de varianza explicada por los efectos fijos, y el R^2 condicional, que representa proporción de varianza explicada por todo el modelo.

En concreto, el R^2 marginal es de **0.702**, lo que significa que aproximadamente el 70.2% de la varianza total en la felicidad se explica exclusivamente por los efectos fijos del modelo, es decir, por las variables explicativas como `freedom`, `support`, `life_exp` o `civil_liberties`. Por otro lado, el R^2 condicional asciende hasta **0.932**, lo que implica que si además se consideran los efectos aleatorios, en este caso las variaciones específicas de cada país con el año, el modelo es capaz de explicar el 93.2% de la varianza total del Happiness Score; mostrando que nuestro modelo explica mucha de la variación de la puntuación felicidad. No obstante, esta gran diferencia de más del 20% entre el R^2 marginal y condicional demuestra que la variabilidad no explicada por los efectos fijos pero capturada por los efectos aleatorios juega un papel clave en la explicación de la felicidad. En conjunto, estos resultados consolidan el poder explicativo del modelo, y que tanto las variables medidas como la modelización de la variabilidad de la felicidad a partir del año contribuyen significativamente a entender las diferencias en los niveles de felicidad.

Como en cualquier modelo estadístico, es esencial verificar que las suposiciones sobre los residuos se cumplen también en el contexto de modelos mixtos. En particular, evaluamos la normalidad de los residuos, la homocedasticidad y proporción de valores atípicos. Para ello, se utilizan gráficos similares a los del modelo clásico, pero adaptados a la estructura jerárquica de los datos. Para validar los supuestos del modelo, se utiliza la función `testResiduals()` del paquete `DHARMa`(Hartig 2024). Esta función genera residuos simulados a partir del modelo ajustado, y los compara con los residuos observados; evaluando tres aspectos fundamentales de los residuos. Primero, la uniformidad, ya que verifica si los residuos simulados siguen una distribución uniforme, lo que sería esperable bajo un modelo bien planteado; basándose en el test de Kolmogorov–Smirnov. Segundo, la dispersión, evaluando la distribución de los residuos comparando la variabilidad observada con la esperada. Y por último, los valores atípicos, porque detecta si hay un número de outliers mayor al esperado bajo el modelo; utilizando una prueba binomial para estimar su proporción.

```
$uniformity

    Asymptotic one-sample Kolmogorov-Smirnov test

data:  simulationOutput$scaledResiduals
D = 0.034426, p-value = 0.06889
alternative hypothesis: two-sided


$dispersion

    DHARMa nonparametric dispersion test via sd of residuals fitted vs.
    simulated

data:  simulationOutput
```

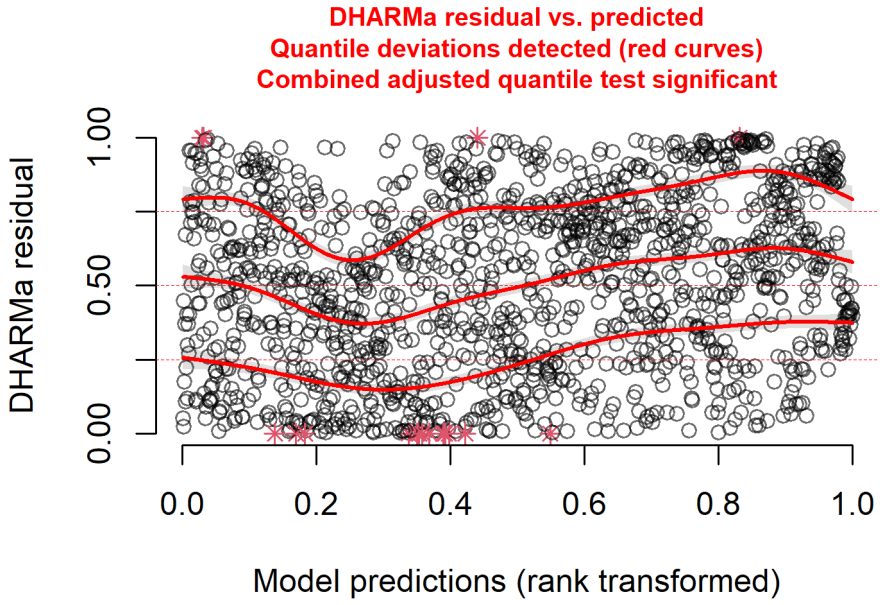

Figura 5.4: Dispersión de los residuos simulados frente a los valores predichos.

```
dispersion = 1.0593, p-value = 0.512
alternative hypothesis: two.sided


$outliers

    DHARMa outlier test based on exact binomial test with approximate
    expectations

data:  simulationOutput
outliers at both margin(s) = 17, observations = 1421, p-value = 0.09827
alternative hypothesis: true probability of success is not equal to 0.007968127
95 percent confidence interval:
 0.006984144 0.019085672
sample estimates:
frequency of outliers (expected: 0.00796812749003984 )
                                                0.01196341
```

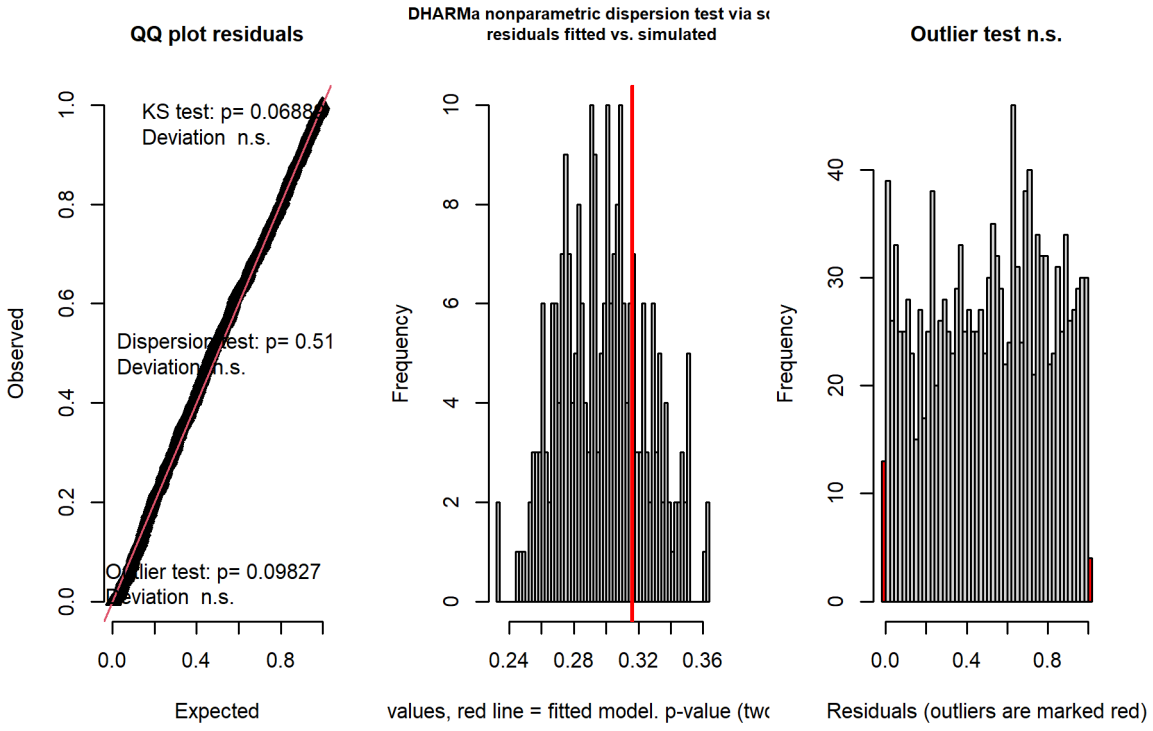

Figura 5.5: Resultados del test formal de uniformidad aplicado a los residuos simulados.

5.4.1 Normalidad

El gráfico QQ-plot de la Figura 5.5 muestra una alineación bastante ajustada de los residuos simulados con la línea teórica, lo que sugiere una distribución aproximadamente normal. Aunque se observan ligeras desviaciones en las colas, el test de Kolmogorov-Smirnov aplicado a los residuos proporciona un p-valor de **0.069**, lo que indica que no se puede rechazar la hipótesis nula de normalidad. Por tanto, los residuos sugieren normalidad y el modelo cumple satisfactoriamente con este supuesto.

5.4.2 Homocedasticidad

En el gráfico de residuos frente a valores ajustados de la Figura 5.4, la nube de puntos es dispersa y no se detecta ninguna tendencia aparente. El test de dispersión de DHARMa refuerza esta conclusión con un p-valor de **0.512**, indicando que no existen evidencias significativas de heterocedasticidad. La varianza de los residuos puede considerarse homogénea, por lo que el modelo cumple también con el supuesto de homocedasticidad.

5.4.3 Outliers y estructura de los residuos

En cuanto a los outliers, el gráfico de residuos simulados frente a predicciones de la Figura 5.5 revela ciertas desviaciones en los cuantiles (indicadas por las líneas rojas), pero si nos fijamos en el test binomial, se estima una proporción de **0.012** de observaciones extremas (17 de 1421), que no difiere significativamente de la esperada por azar (**0.008**). Esto implica que la presencia de valores extremos no es problemática, y como podemos ver, el p-valor es de **0.098**, por lo que el número de outliers no es estadísticamente preocupante.

5.4.4 Conclusión del diagnóstico

El diagnóstico general indica que el modelo cumple de forma razonable con los tres supuestos clave: normalidad, homocedasticidad y proporción esperada de outliers; por lo que se puede considerar estadísticamente fiable para la inferencia y predicción del Happiness Score. Esta validación respalda la solidez del modelo ajustado y su utilidad para explicar las variaciones en la felicidad a partir de las variables seleccionadas.

5.4.5 Predicción del Happiness Score para 2025

Usando el modelo mixto ajustado, se pueden obtener predicciones personalizadas por país; lo que permite construir un ranking proyectado de felicidad para 2025.

Tabla 5.1: Comparación del ranking de felicidad en 2025 y 2024 (Top, España, Bottom)

País	Score 2025	Ranking 2025	Score 2024	Ranking 2024	Sección
Finland	7.754787	1	7.7407	1	Top 10
Denmark	7.605956	2	7.5827	2	Top 10
Iceland	7.523590	3	7.5251	3	Top 10
Netherlands	7.420702	4	7.3194	6	Top 10
Switzerland	7.413425	5	7.0602	9	Top 10
Norway	7.404728	6	7.3017	7	Top 10
Sweden	7.354201	7	7.3441	4	Top 10
Israel	7.257159	8	7.3411	5	Top 10
New Zealand	7.218312	9	7.0292	11	Top 10
Luxembourg	7.204100	10	7.1219	8	Top 10
Spain	6.415798	28	6.4209	35	España
Tanzania	3.579505	139	3.7806	125	Bot 10
Malawi	3.540040	140	3.4210	130	Bot 10
Lesotho	3.519870	141	3.1862	135	Bot 10
Botswana	3.498908	142	3.3834	131	Bot 10
Burundi	3.451707	143	NA	NA	Bot 10
Rwanda	3.421365	144	NA	NA	Bot 10
Zimbabwe	3.278723	145	3.3411	132	Bot 10
Central African Republic	3.231135	146	NA	NA	Bot 10
South Sudan	3.189009	147	NA	NA	Bot 10
Afghanistan	2.539557	148	1.7210	137	Bot 10

Lo primero que podemos observar en la Tabla 5.1 es que los países nórdicos continúan liderando el ranking global del Happiness Score estimado para 2025. Se observa que Finlandia encabeza el ranking con una predicción de 7.75 puntos, consolidando su posición como líder mundial en felicidad global. Le siguen Dinamarca (7.61), Islandia (7.52) y Países Bajos (7.42), todos cerca de su posición en 2024, lo que demuestra cierta estabilidad en su calidad de vida. El top 10 lo completan Suiza, Noruega, Suecia, Israel, Nueva Zelanda y Luxemburgo, todos con puntuaciones superiores a 7.2.

En el extremo opuesto del ranking, los países con los niveles más bajos de felicidad prevista son Afganistán, que ocupa el último puesto con un valor de 2.54, seguido por Sudán del Sur, República Centroafricana y Zimbabue; países que destacan por presentar conflictos, pobreza o inestabilidad política. Todas las puntuaciones están por debajo de 3.6, lo que refleja condiciones estructurales desfavorables que impactan fuertemente de forma negativa en la felicidad.

Si nos fijamos en España, obtiene un Happiness Score previsto de 6.42, situándose en la posición 28; una posición que está por encima de la media. Aunque no alcanza los niveles nórdicos, se

encuentra en el tercio superior del ranking, lo que implica una posición destacada entre los países con mayor felicidad y bienestar. Aunque la puntuación de felicidad predicha para España en 2025 es la misma que para 2024, la posición en el ranking mejora: sube del puesto 35 al 28. Esto implica que otros países han bajado más que España, permitiendo su ascenso en el ranking a pesar de una leve caída en su puntuación; un buen ejemplo de cómo el ranking no depende únicamente de la puntuación, sino también del entorno de comparación entre países.

5.5 Desarrollo del Modelo Lineal Generalizado Mixto (GLMM)

En este apartado buscamos una alternativa al modelo lineal mixto mediante un modelo lineal generalizado mixto (GLMM). Este tipo de modelos permite suavizar el supuesto de normalidad de los residuos, lo que nos sirve de ayuda cuando la variable dependiente presenta asimetría. Como el Happiness Score es una variable positiva y su distribución es ligeramente asimétrica hacia la derecha, utilizaremos una distribución Gamma con enlace logarítmico.

A diferencia del modelo lineal mixto anterior, en este caso haremos una búsqueda automática de modelos válidos combinando variables candidatas en distintos subconjuntos. El principal motivo por el que no utilizamos el mismo modelo que para LMM es que no encontramos ningún GLMM que fuese válido con `year` como efecto aleatorio; por lo que consideramos que utilizar `regional_indicator` también sería coherente para nuestro estudio. Para cada combinación validamos el modelo ajustado aplicando los tests de DHARMa sobre los residuos simulados, exigiendo que se superen todos los test.

El primer modelo válido cuenta con dos efectos fijos: `support` (apoyo social percibido) y `life_exp` (esperanza de vida). En concreto, el modelo es el siguiente:

$$\text{Happiness}_{ij} \sim \text{Gamma}(\mu_{ij}, \theta)$$

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \cdot \text{support}_{ij} + \beta_2 \cdot lifeexp_{ij} + u_{0i} + u_{1i} \cdot regionalindicator_{ij}.$$

donde:

- μ_{ij} es la esperanza del nivel de felicidad del país i en el año j.
- $\beta_0, \beta_1, \beta_2$ son los efectos fijos.
- u_{0i}, u_{1i} son los efectos aleatorios asociados al país i en función de `regional_indicator`.
- El término $\log(\mu_{ij})$ indica que se está usando un link logarítmico.

Aunque hemos explorado cientos de combinaciones de variables explicativas posibles para ajustar un modelo lineal generalizado mixto (GLMM) válido, ninguno de los modelos obtenidos ha devuelto un valor definido para el AIC: en todos los casos, esta métrica ha aparecido como NA (Not Available). El AIC, para que pueda calcularse correctamente, requiere que todos los componentes del modelo hayan sido estimados de forma numéricamente estable. Este problema puede haberse dado debido a que, al contar con todos los países, estamos trabajando con estructuras complejas y variables que pueden estar altamente correlacionadas, lo que hace que

la variabilidad entre regiones lleve a problemas de redundancia o que la combinación de efectos tenga singularidad. Es importante destacar que el hecho de que el AIC sea NA no implica que el modelo no sea válido, sino que no se puede evaluar su calidad frente a otros modelos. Para priorizar la validez estadística del modelo, optamos por aceptar modelos con AIC NA aunque registremos esta limitación. A continuación, se muestra la salida del GLMM:

```
 Family: Gamma  ( log )
Formula:
happiness_score ~ support + life_exp + (1 + regional_indicator |      country)
Data: df_unificado

      AIC       BIC    logLik -2*log(L)  df.resid
       NA        NA        NA        NA      1362

Random effects:

Conditional model:
 Groups  Name                                                Variance
 country (Intercept)                                         1.104e-02
         regional_indicatorCommonwealth of Independent States 6.758e-04
         regional_indicatorEast Asia                         7.271e-04
         regional_indicatorLatin America and Caribbean       2.663e-55
         regional_indicatorMiddle East and North Africa      4.610e-03
         regional_indicatorNorth America and ANZ             8.573e-03
         regional_indicatorSouth Asia                        1.320e-01
         regional_indicatorSoutheast Asia                    3.069e-05
         regional_indicatorSub-Saharan Africa                1.472e-02
         regional_indicatorWestern Europe                    2.554e-03
 Std.Dev.  Corr
 1.051e-01
 2.600e-02 -0.60
 2.696e-02 -0.98  0.73
 5.161e-28 -1.00  0.60  0.98
 6.790e-02  0.93 -0.83 -0.98 -0.93
 9.259e-02  1.00 -0.60 -0.98 -1.00  0.93
 3.633e-01 -1.00  0.60  0.98  1.00 -0.93 -1.00
 5.540e-03 -0.90  0.81  0.95  0.90 -0.98 -0.90  0.90
 1.213e-01  0.98 -0.44 -0.93 -0.98  0.85  0.98 -0.98 -0.81
 5.054e-02  0.72  0.13 -0.58 -0.72  0.42  0.72 -0.71 -0.39  0.84
Number of obs: 1421, groups:  country, 148

Dispersion estimate for Gamma family (sigma^2): 0.00503
```

```
Conditional model:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.4210831  0.0389971   36.44  < 2e-16 ***
support     0.2141082  0.0201328   10.63  < 2e-16 ***
life_exp    0.0020772  0.0004669    4.45 8.63e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Entre los efectos fijos, observamos que tanto `support` como `life_exp` tienen efectos positivos y altamente significativos sobre el nivel de felicidad (**<0.001**). El coeficiente de `support` es de **0.2141**, lo que implica que, manteniendo constantes el resto de variables, un incremento de una décima en el nivel de apoyo social está asociado con un incremento de un 2.2% (exp(0.02141) = 1.022) en el valor esperado del Happiness Score; consolidando la idea de que aumento del apoyo social está asociado al aumento de la felicidad. Por su parte, el coeficiente de `life_exp` es de **0.00207**, lo que significa que por cada año que aumente la esperanza de vida, la felicidad aumenta en torno a un 0.2%; teniendo un impacto menor que el apoyo social pero demostrando que la longevidad y una buena sanidad están positivamente relacionadas con la felicidad. En cuanto a los efectos aleatorios, observamos que la varianza del intercepto entre países (**0.01104**) nos indica que existen todavía diferencias entre países que afectan al nivel de felicidad y que no son explicadas por los efectos fijos del modelo. Si observamos el resto de varianzas, vemos que hay algunas cercanas a 0, como es el caso de Sudamérica; lo cual explica el motivo por el que hemos obtenido un AIC de NA: hay valores cercanos a 0 que hacen que la matriz de varianzas-covarianzas sea singular y no tenga estabilidad numérica. Por último, el modelo presenta un estimador de dispersión (σ^2) de **0.00503**, lo que indica una baja variabilidad residual y, por tanto, un buen ajuste. En general, este modelo GLMM permite plasmar de forma eficaz la relación entre felicidad y otros factores como el apoyo social y la esperanza de vida, teniendo en cuenta variabilidad específica entre regiones.

Se ha calculado el R^2 marginal para el modelo GLMM, obteniendo un valor de **0.387**, lo que indica que los efectos fijos del modelo explican aproximadamente el 38.7% de la varianza del Happiness Score. Sin embargo, no ha sido posible calcular el R^2 condicional debido a que hay componentes de los efectos aleatorios que tienen varianzas cercanas a cero (como es el caso de Sudamérica), lo que genera problemas de singularidad. Este suceso, si bien no nos permite ver del todo la calidad del modelo, no compromete su validez en términos de significancia y ajuste, pero limita la interpretación del componente aleatorio. Al igual que hicimos antes, vamos a comprobar que este modelo es válido para poder hacer predicciones.

```
$uniformity

    Asymptotic one-sample Kolmogorov-Smirnov test

data:  simulationOutput$scaledResiduals
```

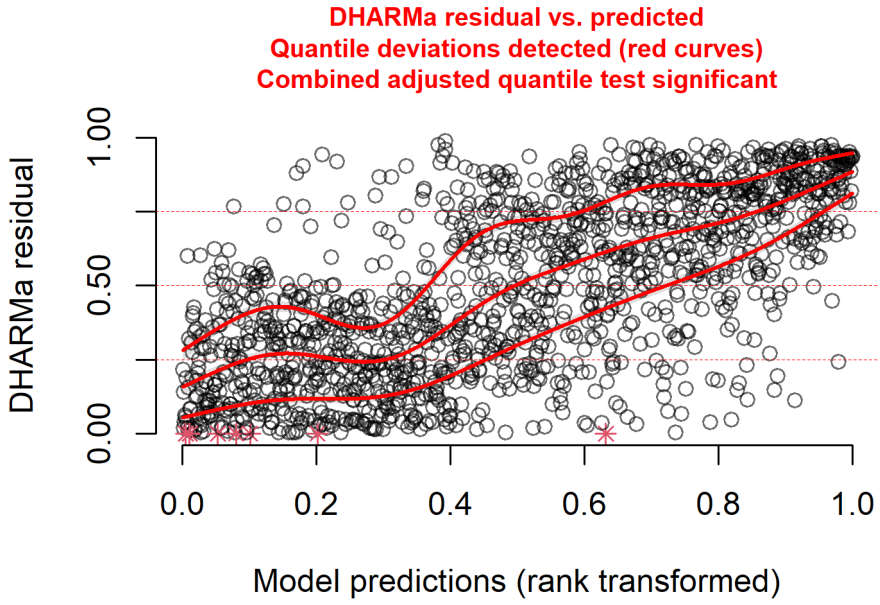

Figura 5.6: Dispersión de los residuos simulados frente a los valores predichos.

```
D = 0.034238, p-value = 0.07148
alternative hypothesis: two-sided


$dispersion

    DHARMa nonparametric dispersion test via sd of residuals fitted vs.
    simulated

data:  simulationOutput
dispersion = 0.81883, p-value = 0.176
alternative hypothesis: two.sided


$outliers

    DHARMa outlier test based on exact binomial test with approximate
    expectations

data:  simulationOutput
outliers at both margin(s) = 7, observations = 1421, p-value = 0.2325
```

```
alternative hypothesis: true probability of success is not equal to 0.007968127
95 percent confidence interval:
 0.001982776 0.010123174
sample estimates:
frequency of outliers (expected: 0.00796812749003984 )
                                             0.004926108
```

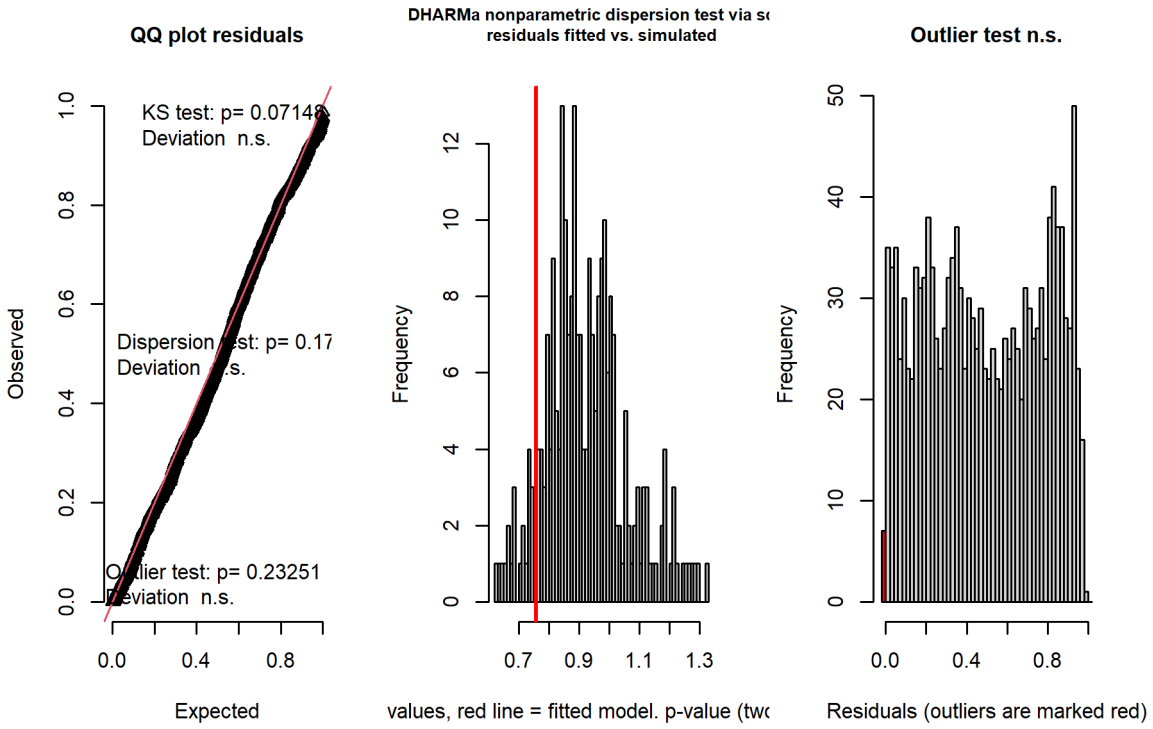

Figura 5.7: Resultados del test formal de uniformidad aplicado a los residuos simulados.

5.5.1 Normalidad de residuos

El gráfico QQ-plot de la Figura 5.7 muestra una alineación bastante razonable con la línea diagonal, sin desviaciones aparentes. El test de Kolmogorov-Smirnov nos da un p-valor de **0.071**, lo que indica que no podemos rechazar la hipótesis de uniformidad, y, por tanto, cumple el supuesto de normalidad.

5.5.2 Homocedasticidad

El gráfico de residuos frente a predicciones de la Figura 5.6 no muestra ningún tipo de tendencia aparente en la dispersión, por lo que gráficamente no parece haber evidencia de heterocedasticidad. En el test de dispersión obtenemos un p-valor de **0.176**, confirmando que no hay evidencias para rechazar la hipótesis de homocedasticidad; cumpliendo con el supuesto.

5.5.3 Outliers y estructura de los residuos

El gráfico de residuos frente a valores ajustados de la Figura 5.7 muestra que hay una parte muy pequeña de observaciones que se desvía del comportamiento esperado, y numéricamente a través del test binomial, que detecta 7 outliers en un total de 1421 observaciones (porporción de **0.005**), vemos que el p-valor obtenido es de **0.233**, lo que indica que la proporción de valores atípicos no significativamente de la proporción esperada; por lo que no hay evidencias de que los outliers estén afectando de manera preocupante al modelo.

5.5.4 Conclusión del modelo GLMM

Este modelo GLMM, a pesar de tener ciertos problemas de singularidad, ha demostrado a través de su diagnóstico que tiene un ajuste razonable y estadísticamente válido. La interpretación de los efectos fijos confirma que tanto el apoyo social como la esperanza de vida son variables explicativas que tienen un impacto positivo en la felicidad, lo cual es coherente con la teoría y con los resultados del modelo LMM. El modelo GLMM ofrece una alternativa al LMM cuando queremos modelar la variable objetivo sin asumir distribución normal, y demuestra ser un método fiable a la hora de querer estimar la felicidad global.

5.5.5 Predicción del Happiness Score para 2025

Al igual que en el modelo lineal mixto, utilizaremos nuestro modelo GLMM para estimar la felicidad en 2025.

Tabla 5.2: Comparación del ranking de felicidad en 2025 (LMM y GLMM) y 2024

País	Score 2025 (GLMM)	Ranking 2025 (GLMM)	Score 2025 (LMM)	Ranking 2025 (LMM)	Score 2024	Ranking 2024	Sección
Finland	7.625442	1	7.754787	1	7.7407	1	Top 10
Denmark	7.537479	2	7.605956	2	7.5827	2	Top 10

Tabla 5.2: Comparación del ranking de felicidad en 2025 (LMM y GLMM) y 2024

País	Score 2025 (GLMM)	Ranking 2025 (GLMM)	Score 2025 (LMM)	Ranking 2025 (LMM)	Score 2024	Ranking 2024	Sección
Iceland	7.483858	3	7.523590	3	7.5251	3	Top 10
Norway	7.417576	4	7.404728	6	7.3017	7	Top 10
Switzerland	7.410418	5	7.413425	5	7.0602	9	Top 10
Netherlands	7.367901	6	7.420702	4	7.3194	6	Top 10
Sweden	7.306411	7	7.354201	7	7.3441	4	Top 10
New Zealand	7.232209	8	7.218312	9	7.0292	11	Top 10
Israel	7.229333	9	7.257159	8	7.3411	5	Top 10
Australia	7.192397	10	7.177458	11	7.0569	10	Top 10
Spain	6.386658	31	6.415798	28	6.4209	35	España
Malawi	3.737572	139	3.540040	140	3.4210	130	Últimos 10
Yemen	3.706376	140	3.598572	138	3.5610	127	Últimos 10
Botswana	3.653473	141	3.498908	142	3.3834	131	Últimos 10
Zimbabwe	3.616604	142	3.278723	145	3.3411	132	Últimos 10
Tanzania	3.575379	143	3.579505	139	3.7806	125	Últimos 10
Rwanda	3.417567	144	3.421365	144	NA	NA	Últimos 10
South Sudan	3.303808	145	3.189009	147	NA	NA	Últimos 10
Burundi	3.297589	146	3.451707	143	NA	NA	Últimos 10
Central African Republic	3.228274	147	3.231135	146	NA	NA	Últimos 10

Tabla 5.2: Comparación del ranking de felicidad en 2025 (LMM y GLMM) y 2024

País	Score 2025 (GLMM)	Ranking 2025 (GLMM)	Score 2025 (LMM)	Ranking 2025 (LMM)	Score 2024	Ranking 2024	Sección
Afghanistan	2.864293	148	2.539557	148	1.7210	137	Últimos 10

Los resultados de la predicción de la Tabla 5.2 muestran que los países nórdicos siguen liderando el ranking de felicidad mundial: Finlandia encabeza la lista con una puntuación estimada de 7.63, seguida por Dinamarca (7.54), Islandia (7.48) y Noruega (7.42). Otros países del top 10 incluyen a Suiza, Países Bajos, Suecia, Nueva Zelanda, Israel y Australia, todos con puntuaciones por encima de 7.1; contando con prácticamente los mismos países que el modelo LMM, aunque difieren en alguna posición.

En contraste, los países con las puntuaciones más bajas en felicidad predicha para 2025 son Afganistán (2.86), República Centroafricana (3.23), Burundi (3.30), Sudán del Sur (3.30), y Ruanda (3.42). Al igual que antes, estos países se caracterizan por tener conflictos, inestabilidad política y pobreza, lo que afecta de forma negativa a su felicidad.

En cuanto a España, se predice un Happiness Score de 6.39 para 2025, lo que la sitúa en la posición 31 del ranking mundial. Esta puntuación representa una ligera reducción respecto al año anterior (6.42), pero supone una mejora en la clasificación, ya que sube del puesto 35 al 31. Esta subida refleja que, aunque la puntuación de España haya sufrido una pequeña bajada, otros países han experimentado caídas mayores, permitiendo que España ascienda en el ranking relativo.

Estos resultados señalan que incluso un modelo con solo dos variables explicativas como lo son el apoyo social y la esperanza de vida puede comprender de forma significativa las diferencias en felicidad entre países. Además, las predicciones realizadas con este GLMM son coherentes con las tendencias observadas, y consolidan al apoyo social y a la esperanza de vida como factores clave de la felicidad a nivel global.

Si hacemos una comparación entre los modelos LMM y GLMM a través de sus predicciones para el ranking de felicidad de 2025, tal como se presenta en la Tabla 5.2, vemos como la mayoría de resultados coinciden, tanto en posiciones superiores como en posiciones inferiores del ranking; demostrando la robustez del modelo mixto para este tipo de datos, independientemente de la familia de distribución asumida. Sin embargo, para decidir qué modelo es mejor utilizaremos como criterio el coeficiente de determinación R^2, ya que es una medida que se utiliza precisamente para evaluar la capacidad explicativa de los modelos. Como hemos visto antes, en el caso del modelo mixto lineal (LMM), el R^2 marginal alcanza un valor de **0.702**, mientras que el R^2 condicional asciende a **0.932**. Por el contrario, el modelo mixto generalizado (GLMM) presenta únicamente un R^2 marginal de **0.387**, sin poder

estimarse el R^2 condicional debido a problemas de singularidad. Esta diferencia muestra que el modelo LMM explica una proporción mucho mayor de la variabilidad observada en el Happiness Score, por lo que concluimos que el modelo LMM es notablemente superior al GLMM en términos de capacidad explicativa y ajuste general a los datos.

A partir de los resultados obtenidos en este capítulo, se ha desarrollado una aplicación interactiva en Shiny que permite explorar dinámicamente los modelos mixtos ajustados sobre los datos de felicidad. La aplicación está diseñada para que el usuario pueda seleccionar distintas combinaciones de efectos fijos y aleatorios, y visualizar tanto los coeficientes estimados como las métricas de calidad del modelo (AIC, R^2 marginal y condicional). Además, en caso de que el modelo generado sea válido, se pueden hacer predicciones personalizadas para el año siguiente, lo que permite observar el comportamiento del modelo bajo distintos escenarios. Esta aplicación refleja de manera interactiva todo el proceso de modelado que se ha ido desarrollando en este capítulo, concediendo al usuario la capacidad de comprobar los diferentes factores que pueden influir en la felicidad y la capacidad de comprender cómo los modelos mixtos se ajustan a datos longitudinales. Dicha aplicación está descrita en el siguiente capítulo.

6 Aplicación Shiny para la modelización de la felicidad

Este capítulo está dedicado a la aplicación interactiva desarrollada con Shiny (Chang et al. 2024) como parte del Trabajo de Fin de Grado. La app permite analizar, modelizar y validar los diferentes factores influyentes de la felicidad a partir de datos longitudinales. Esta aplicación se ha diseñado como una herramienta accesible tanto para usuarios con cierta formación en el ámbito de estadística, como para personas menos formadas que desean comprender mejor la evolución de la felicidad en el mundo.

6.1 Estructura general de la aplicación

El propósito de la aplicación es visualizar la evolución temporal y espacial de variables socioeconómicas y políticas relacionadas con la felicidad, ajustar modelos mixtos lineales (LMM) y generalizados (GLMM) de forma personalizada, validar los modelos ajustados mediante tests estadísticos y gráficos y generar predicciones del Happiness Score para el año 2025 en caso de que el modelo generado sea válido, y permitir al usuario estudiar diferentes combinaciones de variables y configuraciones del modelo. La interfaz de la aplicación está organizada en tres pestañas principales:

6.1.1 Pestaña "Información"

Contiene un resumen del funcionamiento de la app, instrucciones para el uso de cada pestaña, y un enlace al repositorio de GitHub donde se encuentra el código fuente. La interfaz de dicha pestaña la podemos observar en la Figura 6.1.

6.1.2 Pestaña "Descriptiva"

En esta pestaña, el usuario puede seleccionar hasta dos variables numéricas (`happiness_score`, `gdp`, `freedom`, etc.), elegir una o varias regiones del mundo (Western Europe, Sub-Saharan Africa…) y los países cuya evolución de variables a lo largo del tiempo quiera observar, y generar dos tipos de visualización: un gráfico de evolución temporal de las variables y países elegidos, y un mapa mundial para representar el valor de una variable en un año concreto.

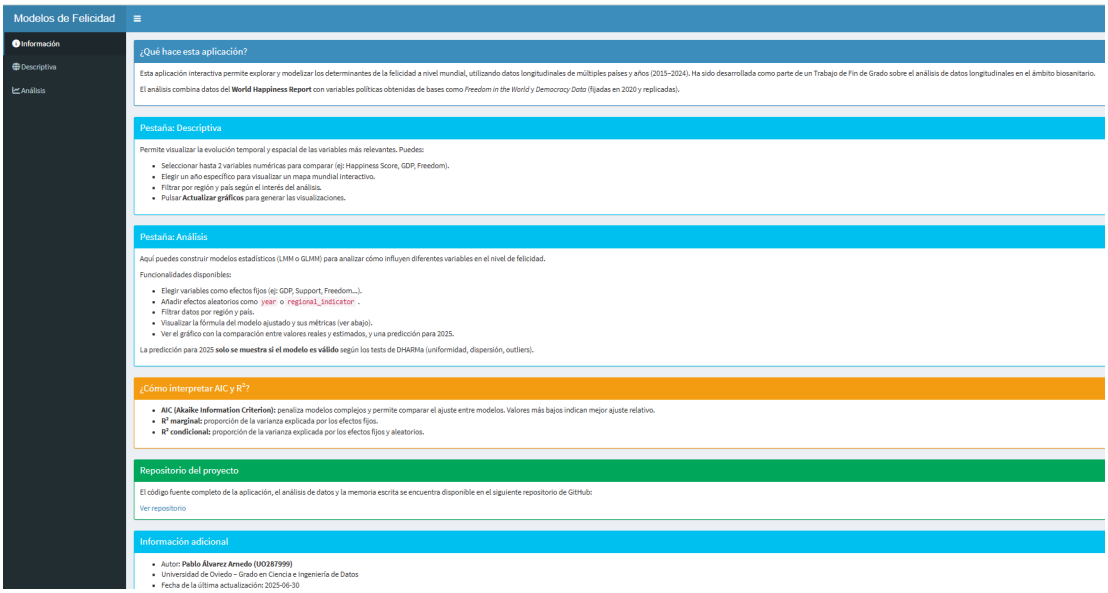

Figura 6.1: Interfaz de la pestaña Información.

De esta manera, hacemos una exploración inicial de los datos de forma visual y podemos identificar posibles tendencias y contrastes regionales. La interfaz de la pestaña "Descriptiva" la encontramos en la Figura 6.2.

6.1.3 Pestaña "Análisis"

Es el punto clave de la aplicación, donde podemos construir y validar modelos estadísticos. Entre las distintas funcionalidades disponibles está la elección de efectos fijos y aleatorios (el usuario puede seleccionar las variables que quiere incluir como efectos fijos (`gdp`, `freedom`, `regime_category`, etc.) y aleatorios (`year` y/o `regional_indicator`)), el filtrado de regiones y la elección de países sobre los que construir y ajustar el modelo y, finalmente, la elección del tipo de modelo: LMM o GLMM. Después de ajustar el modelo, se muestra la fórmula resultante, el resumen estadístico y las métricas de ajuste (AIC, R^2 marginal y condicional). Todo esto lo podemos observar en la Figura 6.3. Luego se realiza la validación del modelo, en la que se muestra la gráfica de residuos vs ajustados y QQ-Plot de residuos, y test de uniformidad, dispersión y outliers. Después de validar el modelo, se produce su diagnóstico: si el modelo es válido, puede hacer predicciones y, en caso contrario, se especifica en qué parte de la validación falla el modelo. Todo esto lo podemos observar en la Figura 6.4. Por último, esta pestaña incluye una gráfica en la que se muestran las observaciones reales y los valores ajustados del modelo. En caso de que el modelo sea válido, se incluyen también las predicciones del modelo para 2025. Esta última parte la observamos en la Figura 6.5. Esta pestaña permite aplicar el conocimiento teórico sobre modelos mixtos explicado en capítulos anteriores, pero en un entorno interactivo que permite diseñar modelos mixtos de forma personalizada.

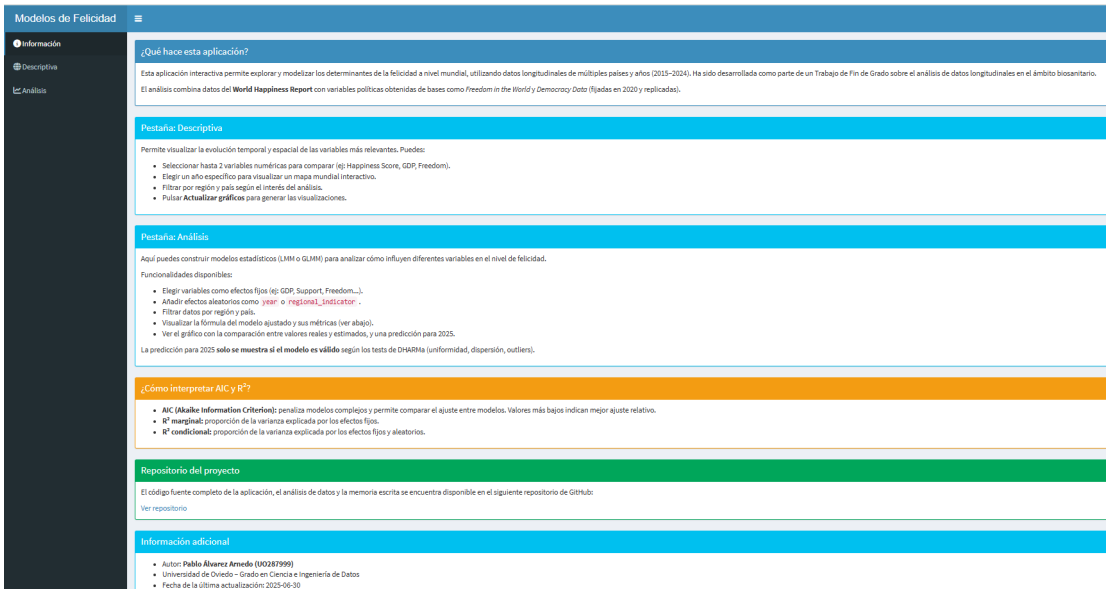

Figura 6.2: Interfaz de la pestaña Descriptiva.

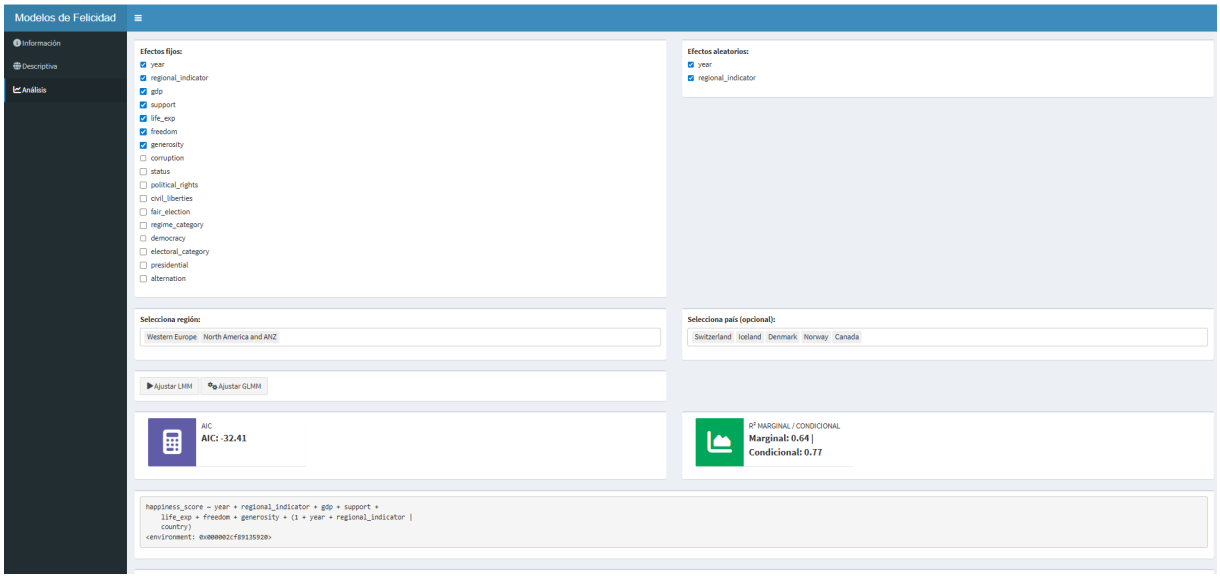

Figura 6.3: Selección de variables y métricas del modelo en la pestaña de análisis.

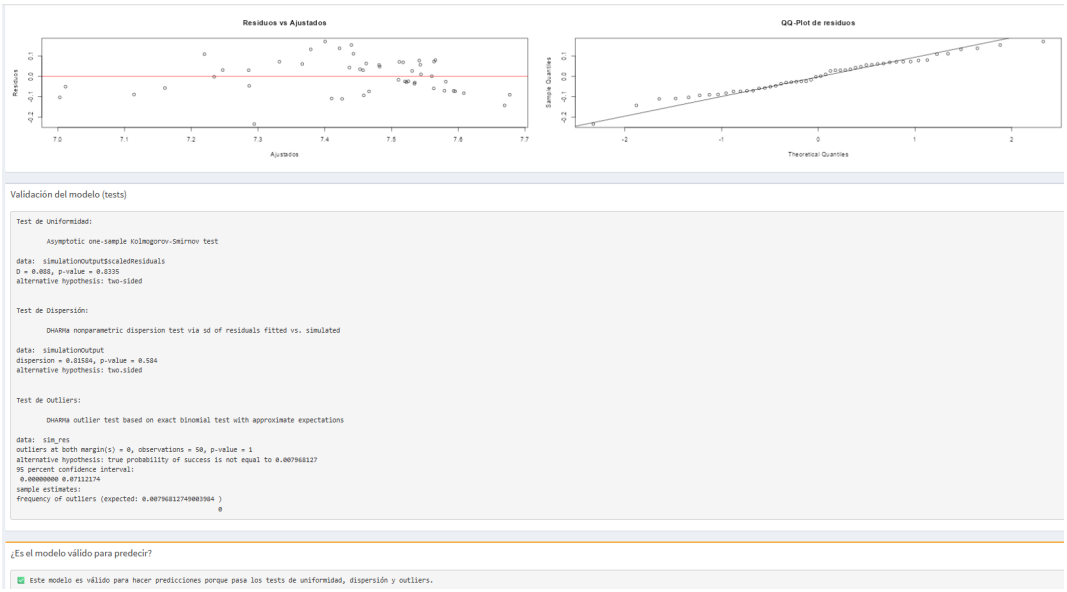

Figura 6.4: Validación del modelo ajustado en la pestaña de análisis.

Figura 6.5: Gráfico de predicciones por país y año en la pestaña de análisis.

6.2 Integración con el análisis longitudinal

La aplicación no analiza los datos para un año concreto, sino que está diseñada para trabajar con datos longitudinales que recogen medidas repetidas de múltiples países a lo largo de los años. Esto permite reflejar la evolución temporal de la felicidad y ajustar modelos que consideren tanto las diferencias entre países como los cambios dentro de cada país a lo largo del tiempo. Al igual que en el capítulo anterior, la aplicación cuenta con una base de datos que contiene todas las variables políticas explicativas, lo que permite ajustar modelos con componentes temporales, regionales, y políticas; reflejando la estructura multinivel de este tipo de datos. Además, la aplicación otorga al usuario la capacidad de construir modelos de forma intuitiva, permitiendo en todo momento seleccionar y deseleccionar los efectos, continentes y países que el usuario considere oportunos.

La interfaz orienta al usuario en todo momento, mostrando en tiempo real la fórmula del modelo ajustado, su resumen estadístico, las métricas de ajuste y los resultados de validación. Esta disposición de la aplicación, desde la selección de variables hasta la validación final del modelo, permite que el usuario comprenda la forma en la que cada variable contribuye al modelo y cuál es el impacto que tiene cada una de ellas en la puntuación de la felicidad.

6.3 Repositorio de GitHub

El desarrollo completo de la aplicación, incluyendo los códigos y datos procesados y el archivo de la app Shiny, se encuentra disponible en un repositorio público de GitHub. Este repositorio simplifica la reproducibilidad del trabajo y funciona como plataforma de difusión y colaboración. Cualquier usuario tiene la capacidad de acceder al repositorio, ejecutar la app localmente y revisar el código fuente. Además, se ha incluido un resumen del repositorio en el que se describe su contenido, el cual incluye no sólo el código relacionado a la aplicación, sino a la memoria del Trabajo de Fin de Grado.

7 Conclusiones y mejoras futuras

En este último capítulo recopilamos los principales éxitos del trabajo, analizando su evolución a lo largo de las diferentes fases, y sugiriendo posibles líneas de mejora. El propósito de este capítulo es determinar el cumplimiento de los objetivos iniciales y valorar, de manera crítica, los resultados obtenidos.

7.1 Resumen y aportaciones realizadas

El trabajo parte de un objetivo claro: estudiar cómo se pueden aplicar los modelos lineales mixtos (LMM) y generalizados mixtos (GLMM) al análisis de datos longitudinales en un problema real. Para ello, se escogió como caso de estudio la evolución del Happiness Score a lo largo del tiempo, incorporando diferentes variables sobre factores sociales, económicos y políticos. A través de este método, podemos enseñar el potencial que tienen este tipo de modelos a la hora de adaptarse a datos longitudinales, en este caso datos donde se quiere capturar la variabilidad temporal y geográfica en un suceso como puede ser la felicidad global.

Esta evolución no solo supuso una mejora progresiva hasta obtener el resultado final, sino también un proceso de aprendizaje continuo: desde el manejo de modelos mixtos en R hasta el desarrollo de una aplicación interactiva en la que poder integrar estos modelos. En general, el proyecto ha acabado en una herramienta robusta y eficaz que permite analizar, modelizar y predecir la felicidad global de manera sencilla y accesible.

El trabajo ha seguido un orden a través de la estructuración en capítulos. En el capítulo 2 se introdujo el concepto de datos longitudinales, señalando sus características y las limitaciones de los métodos de estadística clásica cuando se trabaja con ellos. En el capítulo 3 se presentaron detalladamente los modelos adecuados para datos longitudinales: modelos lineales mixtos (LMM) y modelos lineales generalizados mixtos (GLMM); analizando sus fundamentos matemáticos, componentes, y métodos de estimación y validación. En el capítulo 4 se realizó una exploración, limpieza y mejora de la base de datos World Happiness Report a partir de la integración de dos fuentes externas (Freedom House y Democracy Data) para agregar contexto político a la base de datos. En el capítulo 5 se construyeron diferentes modelos, demostrando que los métodos de estadística clásica no se adaptaron bien a nuestros datos longitudinales, mientras se formularon modelos LMM y GLMM válidos en los que se evaluó la influencia de diferentes variables sobre la felicidad; finalizando el capítulo determinando la capacidad predictiva de estos modelos. Finalmente, en el capítulo 6 se diseñó una aplicación

interactiva en Shiny que permite desarrollar todo este proceso de manera visual y dinámica, explorando y ajustando modelos para luego validarlos y, en caso de ser válidos, generar predicciones.

Desde un enfoque metodológico, el trabajo demuestra cómo aplicar modelos mixtos y generalizados mixtos a un caso real con datos longitudinales, respetando su estructura y dependencia y afrontando diferentes obstáculos como la multicolinealidad o la singularidad en algunos casos. Mirando la aplicación práctica, el trabajo permite estudiar la evolución de la felicidad global de forma flexible, otorgando al usuario la capacidad de personalizar diferentes modelos según las variables y países que considere oportuno; pudiendo estudiar la influencia de ciertos factores como puede ser la esperanza de vida o la percepción de corrupción en la felicidad ciudadana. Observando el aprendizaje obtenido en este trabajo, he adquirido conocimientos en aspectos clave como la programación en R, el uso de librerías como `lme4` o `glmmTMB`, el diseño de aplicaciones Shiny y el uso de plataformas como GitHub para compartir y documentar el código y el avance que iba haciendo a lo largo del trabajo. La app desarrollada no solo representa el resultado final, sino también una demostración del conocimiento adquirido durante el proceso.

Las técnicas desarrolladas en este trabajo no se aplican únicamente en nuestro caso concreto, sino que son fácilmente transferibles a otros conjuntos de datos longitudinales en los que habría que realizar un análisis exploratorio, limpieza e identificación de las características más relevantes como se ha hecho en este trabajo; planteando los modelos pertinentes y razonando la estructura de efectos fijos y aleatorios.

La aplicación Shiny conforma una pieza clave del trabajo, ya que supone la traducción de la teoría de modelos y análisis explicada en los capítulos anteriores a la práctica; facilitando la interpretación de los resultados y ofreciéndole al usuario un entorno en el que poder poner en práctica lo visto hasta ahora. Entre sus principales ventajas destacan la accesibilidad, ya que lo único que se necesita para usar la app es el repositorio de Github; su flexibilidad, ya que permite construir modelos personalizados según elija el usuario; la interactividad, ya que la aplicación permite ver en tiempo real el efecto que tienen las distintas configuraciones del modelo; la validación automática, que permite realizar diagnósticos del modelo de manera rápida y eficaz, indicando en todo momento si el modelo es válido y justificando por qué no lo es; y la fiabilidad de sus predicciones, ya que las estimaciones para el año 2025 sólo se producen si el modelo es estadísticamente válido.

7.2 Limitaciones y posibles mejoras

Pese a los resultados obtenidos, el trabajo también presenta algunas limitaciones, ya que la poca variabilidad que se produce en ocasiones entre países según el efecto aleatorio hace en algunos momentos no se pueda calcular el AIC debido a problemas de singularidad. Además, el modelo predictivo se limita a una predicción puntual para 2025, sin intervalos de confianza en los que se muestre la incertidumbre de dichas predicciones.

Existen varias mejoras que podrían implementarse en el futuro para ampliar la capacidad del trabajo, como la ampliación de diferentes fuentes de datos que nos permitan utilizar las variables como longitudinales en vez de fijas, la incorporación de modelos más complejos o la exportación de resultados desde la app.

En definitiva, en este trabajo se ha explicado cómo utilizar métodos estadísticos avanzados, los LMM y los GLMM, a un problema complejo y relevante como la felicidad global. La combinación de estos modelos con datos longitudinales y la aplicación interactiva han resultado en una herramienta en la que no sólo se ha llevado a la práctica la teoría presentada en este trabajo, sino que también se han visto reflejados los conocimientos adquiridos durante el proceso. Aunque existen claras líneas de mejora, los resultados obtenidos suponen una base sólida a partir de la cual se puede extender el trabajo para poder estudiar en mayor profundidad, desde la estadística, la evolución y análisis de la felicidad global.

Referencias

Bates, Douglas, Martin Mächler, Ben Bolker, y Steve Walker. 2015. «Fitting Linear Mixed-Effects Models Using lme4». *Journal of Statistical Software* 67 (1): 1-48. https://doi.org/10.18637/jss.v067.i01.

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, y Jonathan McPherson. 2024. *shiny: Web Application Framework for R*. https://CRAN.R-project.org/package=shiny.

Faraway, Julian J. 2006. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Boca Raton, FL: Chapman & Hall/CRC.

Freedom House. 2024. «Freedom in the World». https://freedomhouse.org/report/freedom-world.

Hartig, Florian. 2024. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. https://CRAN.R-project.org/package=DHARMa.

Helliwell, John F., Haifang Huang, Shun Wang, y Max Norton. 2023. «World Happiness Report 2023». Sustainable Development Solutions Network. https://worldhappiness.report/ed/2023/.

Helliwell, John F., Richard Layard, Jeffrey D. Sachs, Jan-Emmanuel De Neve, Lara B. Aknin, y Shun Wang, eds. 2024. *World Happiness Report 2024*. Oxford, UK: Wellbeing Research Centre, University of Oxford. https://worldhappiness.report/ed/2024/.

Hernández-Barrera, Francisco. 2024. «Modelos mixtos con R». 2024. https://fhernanb.github.io/libro_modelos_mixtos/.

Inglehart, Ronald, Roberto Foa, Christopher Peterson, y Christian Welzel. 2008. «Development, Freedom, and Rising Happiness: A Global Perspective (1981–2007)». *Perspectives on Psychological Science* 3 (4): 264-85. https://doi.org/10.1111/j.1745-6924.2008.00078.x.

M, McGillycuddy, Warton D. I., Popovic G, y Bolker B. M. 2025. «Parsimoniously Fitting Large Multivariate Random Effects in glmmTMB». *Journal of Statistical Software* 112 (1): 1-19. https://doi.org/10.18637/jss.v112.i01.

McCulloch, Charles E., Shayle R. Searle, y John M. Neuhaus. 2008. *Generalized, Linear, and Mixed Models*. 2nd ed. Hoboken, NJ: John Wiley & Sons.

Pinheiro, Jose, Douglas Bates, y R Core Team. 2024. *nlme: Linear and Nonlinear Mixed Effects Models*. https://CRAN.R-project.org/package=nlme.

R Core Team. 2024a. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

———. 2024b. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

R4DS Online Learning Community. 2024. «TidyTuesday: 2024-11-05 dataset». https://github.com/rfordatascience/tidytuesday/blob/main/data/2024/2024-11-05/readme.md.

Roback, Paul, y Julie Legler. 2021. *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R*. Chapman & Hall/CRC. https://bookdown.org/roback/bookdown-BeyondMLR/.

Subirana, Isaac. 2020. «Curso de datos longitudinales». 2020. https://bookdown.org/isubirana/longitudinal_data_analyses/.