Documento base

Pablo Álvarez Arnedo

2026-02-11

Tabla de contenidos

1 Introducción

```
1 + 1
```

```
[1] 2
```

2 Datos longitudinales

2.1 Datos con medidas repetidas

Los **datos longitudinales** son aquellos que obtenemos al realizar distintas medidas a un individuo (personas, regiones, células, etc.). Dichas medidas se pueden observar repetidamente a lo largo del tiempo (análisis temporal), como el ingreso 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 la causa más usual de medidas repetidas es el tiempo, haremos referencia a este caso en concreto, aunque los otros dos también serían aplicables. Por esto, a los datos longitudinales también se les conoce como medidas repetidas.

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 que nos sería difícil descubrir 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 o tratamientos. En el ámbito biosanitario, los datos longitudinales son fundamentales para investigar la progresión de enfermedades, la efectividad 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, fallan al 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 (e.g., pacientes, regiones). Todo ello significa que cada unidad tiene varias observaciones en diferentes momentos temporales; haciendo que los datos longitudinales adopten una estructura jerárquica.

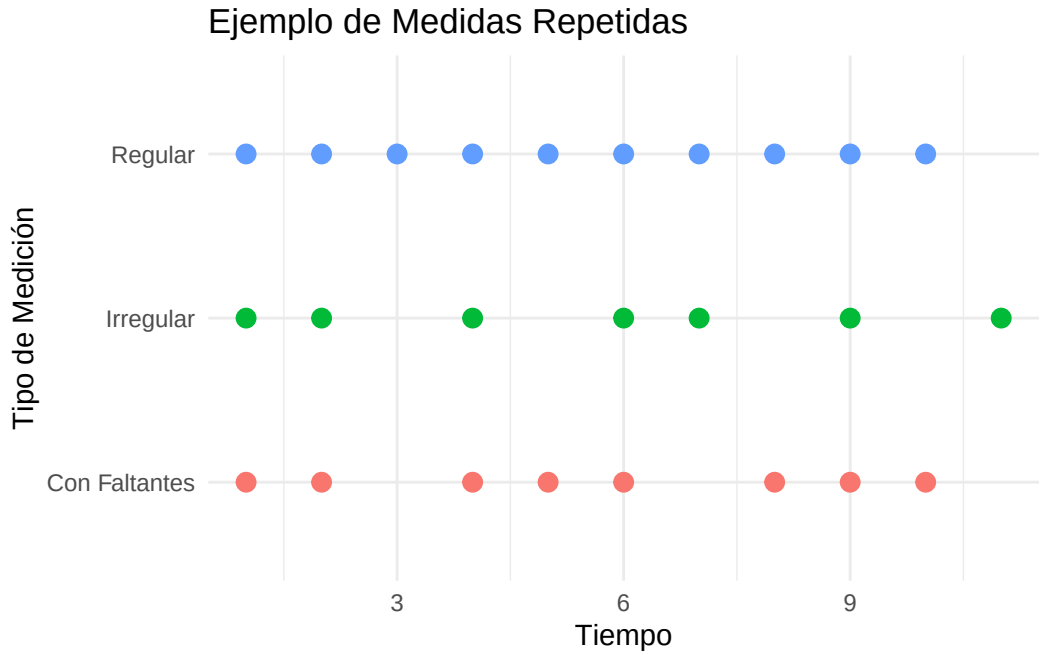
Esta estructura nos lleva a asumir una de las claves en todo este proceso, la dependencia entre las observaciones, la cual nos indica que las mediciones dentro de la misma unidad tienden a estar correlacionadas. También tenemos que destacar las distintas variables que definen a dichos datos, que suelen clasificarse según diferentes propiedades. Como la mayoría de medidas se realizan en distintos instantes de tiempo, es normal que su valor varíe 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 la edad) 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 abordar preguntas específicas 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 o fluctuaciones 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 regulan la evolución de dicha variable; en cuyo caso deberíamos de estudiar cómo dichos factores interactúan con 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 científicos. Por ejemplo, los datos longitudinales tienen una gran importancia en el ámbito biosanitario, como puede ser en estudios donde hay medidas repetidas de presión arterial en un grupo de pacientes durante un tratamiento que nos permiten monitorear la salud de los pacientes para poder evaluar la efectividad del tratamiento e identificar posibles efectos secundarios. No obstante, este tipo de datos también tiene su relevancia en otras áreas como la educación; por ejemplo, la evaluación de los puntajes de un estudiante a lo largo de varios exámenes anuales podría identificar áreas de mejora por parte del alumnado o algunas estrategias pedagógicas que se puedan implementar en la docencia. Otra de las áreas en la que los datos longitudinales juegan un papel clave es la alimentación mediante el estudio de diferentes dietas a diferentes grupos de la población a lo largo del tiempo a través de medidas tales como la actividad física, peso corporal, nivel de colesterol, etc. y cómo estas rutinas aportan ciertos beneficios o riesgos a la salud de los individuos. En otros ámbitos como en el marketing también nos encontramos con casos en los que se utilizan datos longitudinales, como son encuestas de opinión realizadas periódicamente a las mismas personas que pueden ser de gran utilidad 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, el acceso a servicios de salud y la cohesión social influyen en el bienestar de la población. Estos estudios pueden ser fundamentales para que los gobiernos diseñen políticas públicas 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, la migración o el desarrollo humano en diferentes regiones del mundo, proporcionando información valiosa para la toma de decisiones a nivel global.

A pesar de su gran utilidad, los datos longitudinales presentan varias complicaciones adicionales. En primer lugar, aunque las mediciones suelen realizarse en intervalos de tiempo predefinidos, no siempre disponemos de todas las observaciones esperadas debido a la presencia de valores faltantes. Esto puede ocurrir por razones como 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 necesariamente son medidos en los mismos instantes de tiempo, por lo que no siempre tenemos el mismo número de mediciones repetidas por individuo, lo que lleva a una estructura desigual en los datos que debe ser abordada con técnicas adecuadas. Estas dificultades pueden generar desafíos en el modelado y en la comparación de trayectorias individuales, por lo que es fundamental aplicar estrategias estadísticas como imputación de valores faltantes, modelado con efectos aleatorios o técnicas específicas 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 heterogeneidad de las observaciones. Esto se puede apreciar en el siguiente ejemplo:



Como podemos apreciar en la gráfica, 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 mediciones 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:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

donde:

- Y_i representa la variable dependiente (respuesta).
- X_i 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_i 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 confiables, deben cumplirse ciertos **supuestos** fundamentales:

1. **Linealidad:** La relación entre la variable independiente X y la dependiente Y debe ser lineal. Esto 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 ($\varepsilon_i \sim N(0, \sigma^2)$). Esto es especialmente importante para hacer inferencias estadísticas 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 estos supuestos se cumplen, la regresión lineal simple proporciona **estimaciones insesgadas** de los coeficientes y permite hacer inferencia sobre la relación entre X y Y mediante pruebas de hipótesis y construcción de intervalos de confianza.

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

La estadística clásica (e.g., 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 mediciones repetidas de la misma unidad a lo largo del tiempo. Los longitudinales presentan características específicas que requieren enfoques 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 mediciones repetidas pueden estar influenciadas por factores externos o por variables no observadas, lo que genera una relación entre los errores que los modelos tradicionales 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 homogénea, lo que 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 convencionales, como la regresión lineal simple, no sea adecuado para el análisis de datos longitudinales. En su lugar, es necesario recurrir a enfoques específicos, como los modelos lineales mixtos, que permiten modelar tanto los efectos fijos como los efectos aleatorios para capturar adecuadamente la variabilidad y dependencia inherente a 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 (*psid*). Vamos a utilizar un modelo regresión lineal simple para modelar los ingresos en función del tiempo, ignorando la correlación entre mediciones.

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, en caso afirmativo, 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.

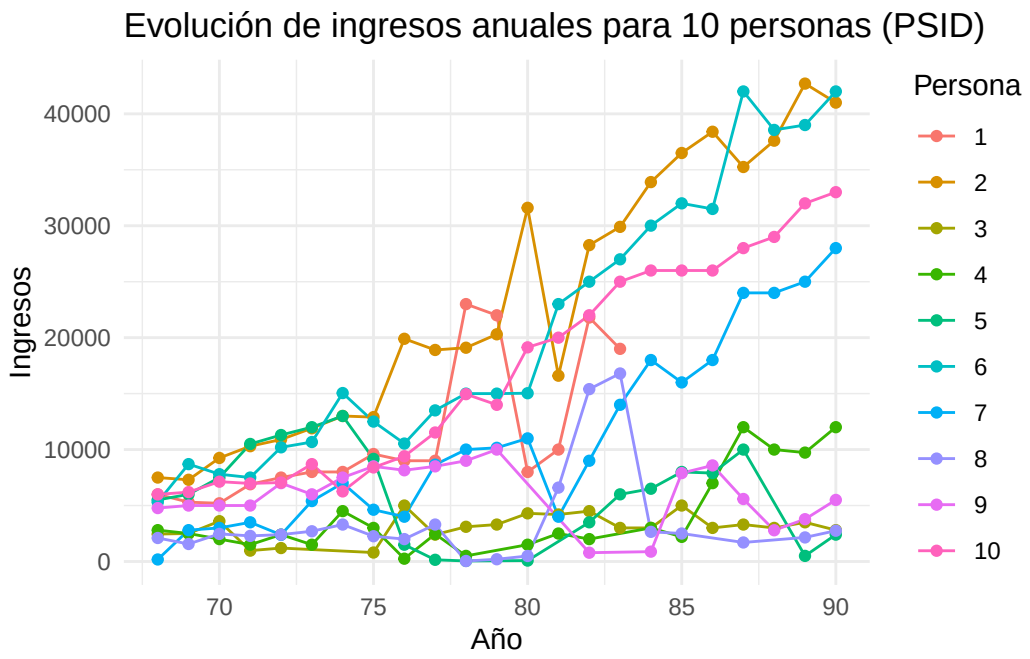


Figura 1. Evolución de los ingresos anuales de 10 personas a lo largo del tiempo.

La figura 1 muestra la evolución de los ingresos anuales para diferentes personas a lo largo del tiempo, en el que cada línea representa a una persona.

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. 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 básicos de independencia de las observaciones, fundamentales para modelos clásicos como la regresión lineal simple.

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 muestra el ajuste de la regresión lineal simple aplicada a los datos.

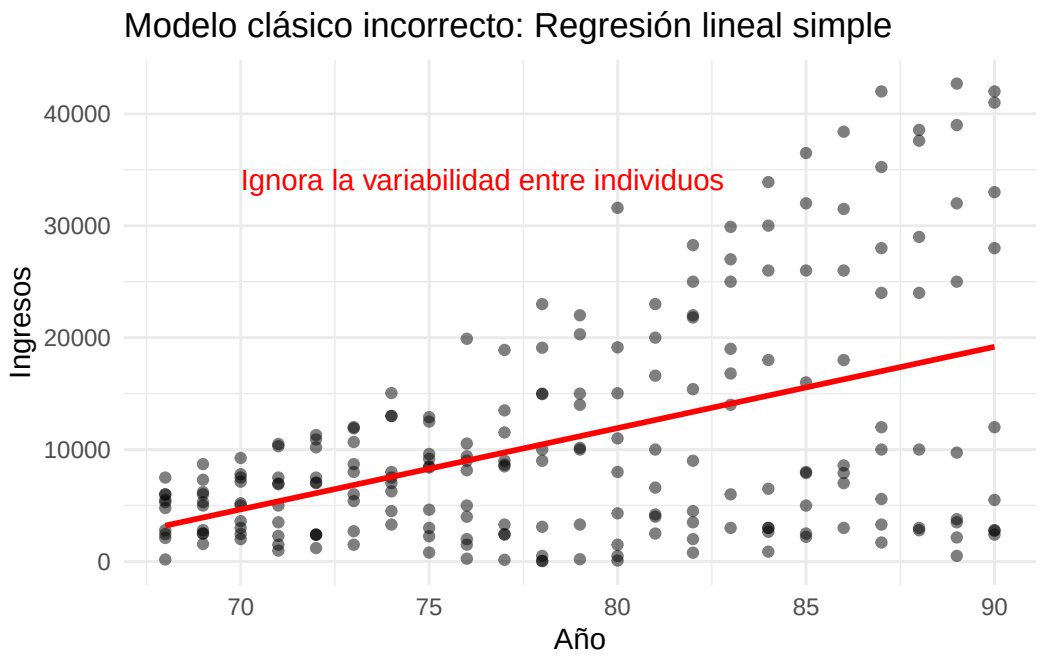


Figura 2. Ajuste de un modelo de regresión lineal simple a los datos observados. La línea roja representa la predicción para cada año, con una única pendiente que asume que todos los individuos comparten la misma relación entre ingresos y tiempo.

Este gráfico muestra cómo la regresión lineal simple aplicada a estos datos genera una representación distorsionada, ignorando por completo la correlación de los datos longitudinales; dando lugar a un mal ajuste y a resultados estadísticos inapropiados 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 tengan media cero, la no correlación y la homocedasticidad.

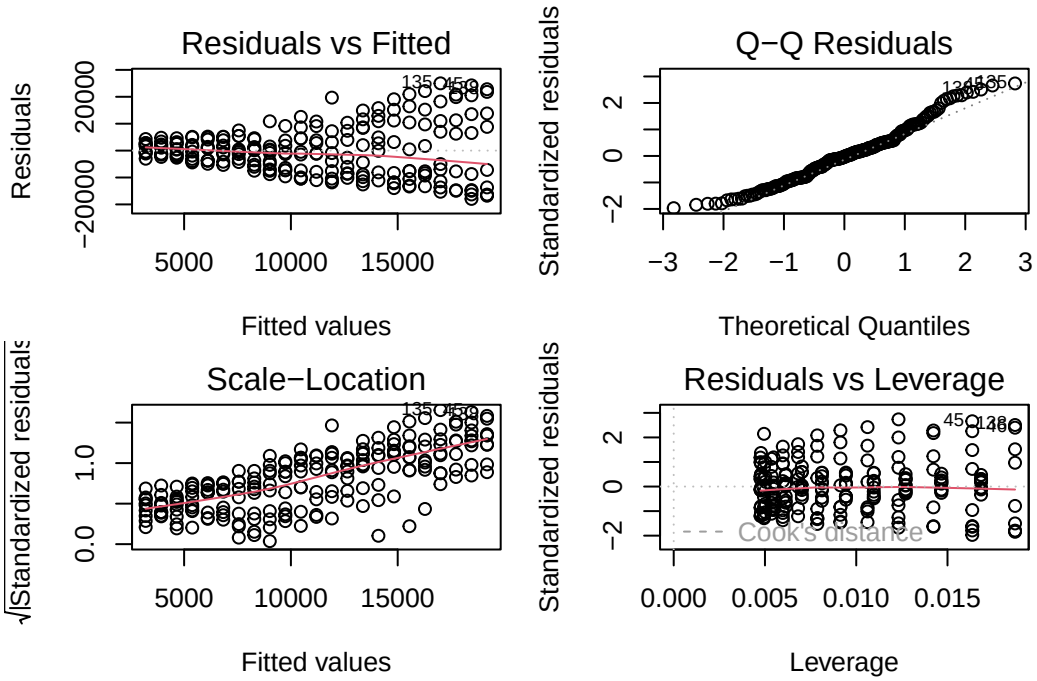


Figura 3. Gráficas de los residuos del modelo.

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

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

Si calculamos la media de los residuos del modelo, comprobamos que la media es **0**, pero esta no es una forma correcta de anlizar la media cero ya que esto no significa que la suposición de media cero se cumpla en todas partes del rango de los valores ajustados. Para hacer un correcto análisis, nos vamos a fijar en la primera gráfica de la figura 3: Residuals vs Fitted. Teóricamente, para que los residuos tengan media cero, deberían de estar uniformemente dispersos alrededor del eje horizontal en $y = 0$. 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, se observa un patrón curvilíneo a medida que aumenta el valor de los valores 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**) 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 3, 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. De hecho, también puede ser el caso paradigmático de normalidad heterocedástica, en la que la varianza depende de la media. 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:

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

Si el p-valor es menor a un umbral significativo (por defecto decimos que es 0.05), se rechaza la hipótesis nula, indicando que los residuos no siguen una distribución normal.

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). 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 4.

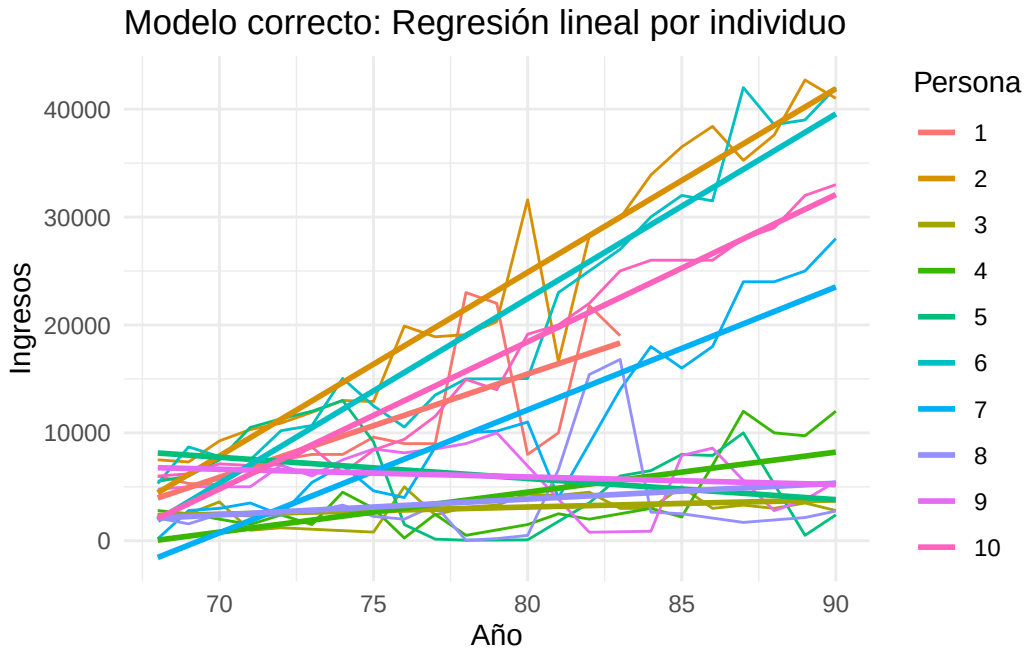


Figura 4. Gráfica del modelo para cada individuo.

En esta gráfica, podemos observar que cada individuo tiene un comportamiento único 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 individuos. Este resultado destaca la heterogeneidad presente 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 hacer inferencias generales sobre la población; además de que en escenarios con un gran número de individuos, esta aproximación no es práctica. Por ello, los **modelos mixtos**, que se explicarán en el siguiente capítulo, emergen como una solución adecuada, ya que combinan los llamados efectos fijos y aleatorios para capturar tanto las tendencias generales de la población como las diferencias específicas entre individuos. Esta aproximación ofrece un equilibrio entre flexibilidad y generalización, respetando las características únicas de los datos longitudinales.

3 Modelos mixtos

En este capítulo, exploraremos los **Modelos Lineales Mixtos (LMM)** y los **Modelos Lineales Generalizados (GLM)**, dos enfoques estadísticos fundamentales para el análisis de datos longitudinales. Veremos cómo los LMM permiten modelar la variabilidad entre individuos mediante la inclusión de efectos aleatorios y fijos, lo que facilita el estudio de la correlación entre observaciones repetidas. Luego, introduciremos los GLM, que extienden la regresión lineal para manejar variables respuesta que no siguen una distribución normal, utilizando funciones de enlace y la familia exponencial. A lo largo del capítulo, revisaremos sus formulaciones matemáticas, sus hipótesis clave y cómo validarlas en la práctica.

Para ilustrar estos modelos, comenzaremos con un ejemplo aplicado al conjunto de datos Orthodont del paquete nlme, donde analizaremos la evolución de la distancia entre los dientes (distance) en función de la edad (age) en diferentes sujetos. Compararemos tres enfoques distintos:

- Modelo con sólo **efectos fijos**: Se asume que todos los sujetos siguen la misma relación.
- Modelo con sólo **efectos aleatorios**: Se permite que cada sujeto tenga su propio valor inicial (intercepto), pero no afecta la pendiente.
- **Modelo mixto**: Se permite que tanto el intercepto como la pendiente varíen entre sujetos.

3.1 Comparación de modelos con efectos fijos, aleatorios y mixtos

La base de datos Orthodont proviene del paquete nlme en R y contiene información sobre el crecimiento dental en niños. 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).

A continuación, ajustaremos y visualizaremos los distintos modelos.

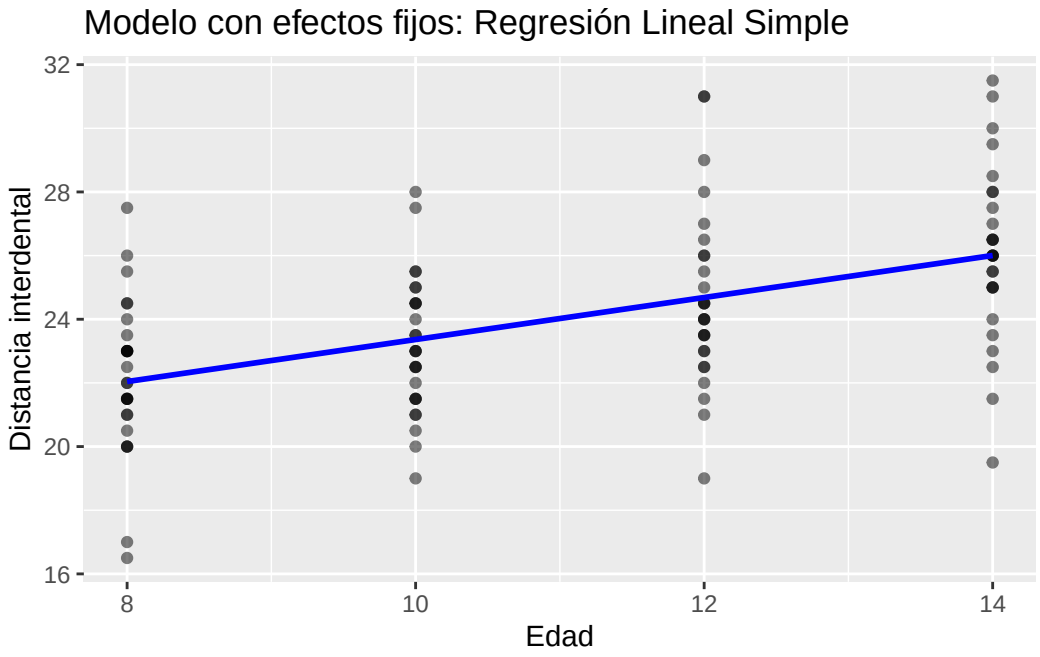
3.1.1 Modelo con efectos fijos

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), pero asumimos que todas las observaciones son independientes e ignoramos la estructura jerárquica del estudio (mediciones 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, β_0 es la intersección común a todos los sujetos, β_1 es la pendiente (cómo cambia la distancia con la edad), y ϵ_i es el error aleatorio.

```
`geom_smooth()` using formula = 'y ~ x'
```



Este modelo considera únicamente la edad (age) como predictor de la distancia (distance) y no tiene en cuenta que los datos son mediciones repetidas de los mismos individuos, lo que puede llevar a errores de estimación debido a la correlación entre observaciones de un mismo sujeto. Como podemos comprobar a través de este ejemplo, si se ignora la estructura jerárquica, podríamos obtener estimaciones erróneas de la variabilidad en la población; obteniendo un coficiente de determinación R^2 bajísimo (**0.256**).

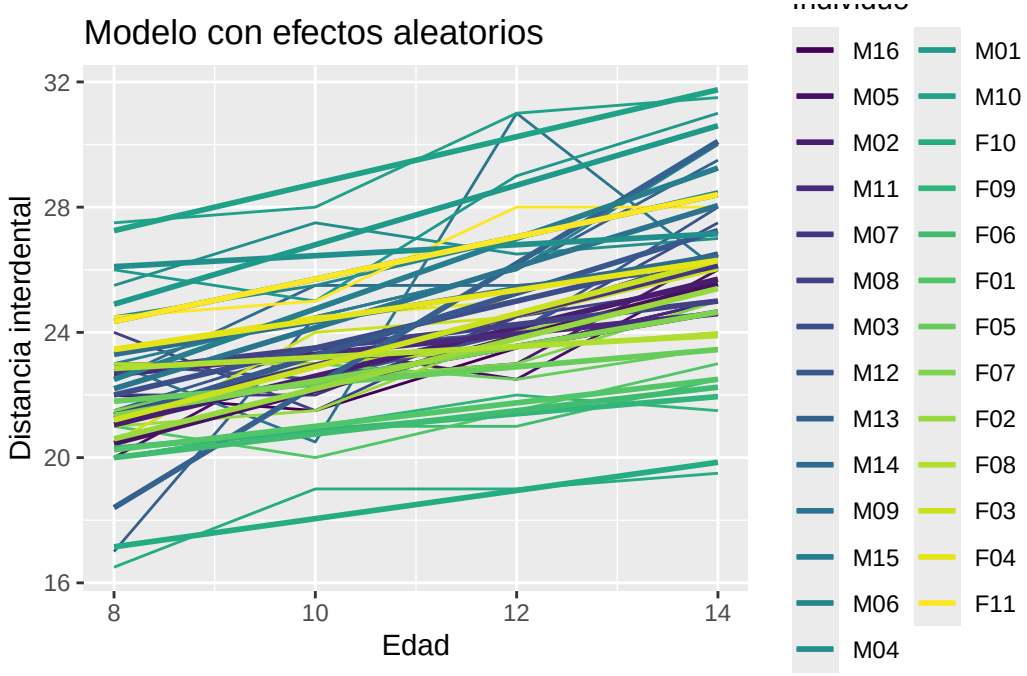
3.1.2 Modelo con efectos aleatorios

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

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

Ahora $distance_{ij}$ es la distancia para el individuo i en la observación j. u_i representa el efecto aleatorio de cada sujeto (intersección específica), β_1 es la pendiente común, y ϵ_{ij} es el error.

```
`geom_smooth()` using formula = 'y ~ x'
```



Ahora tenemos un término indica que cada individuo (Subject) tiene su propia intersección aleatoria, permitiendo que la relación entre la distancia y la edad varíe entre individuos en lugar de asumir una única intersección fija para todos. Esto significa que algunos sujetos pueden tener valores iniciales más altos o más bajos de distance sin que eso afecte la tendencia general de la población. La diferencia crucial 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 cada sujeto tiene un punto de partida diferente en distance, pero que todavía hay una parte de la variabilidad del modelo que no se explica por los efectos fijos ni por las diferencias entre sujetos.

Este modelo permite que cada niño tenga su propio intercepto aleatorio, modelando mejor la variabilidad individual.

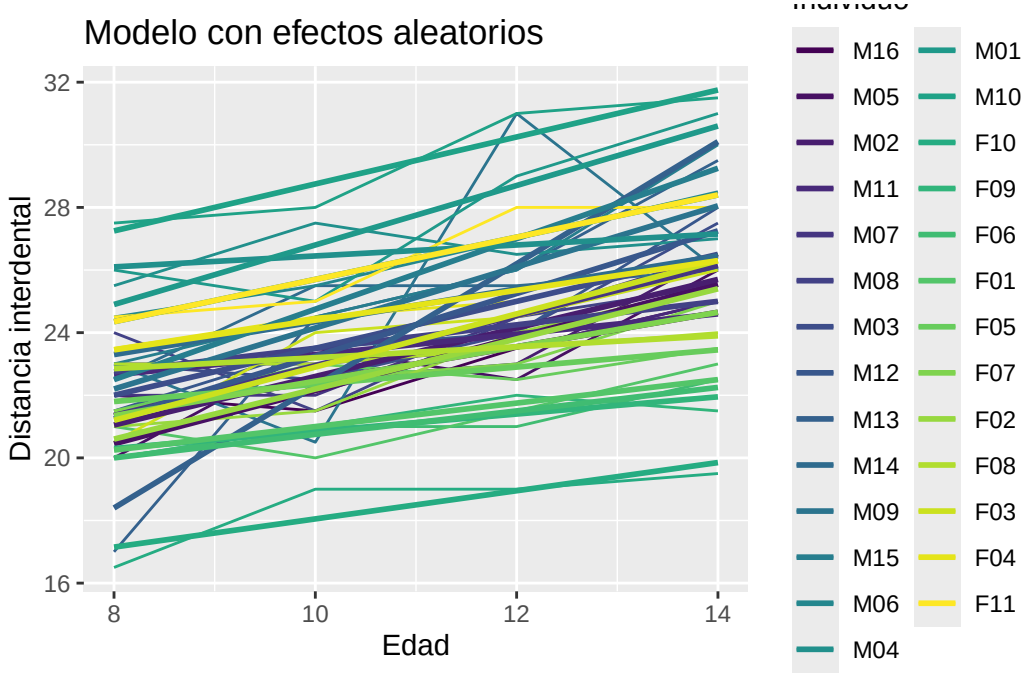
3.1.3 Modelo mixto

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 la intersección específica de cada sujeto, y v_i permite que la pendiente también varíe por individuo.

```
`geom_smooth()` using formula = 'y ~ x'
```



Ahora no solo permitimos una intersección aleatoria, sino que también permitimos que la pendiente (efecto de la edad) varíe entre sujetos; es decir, en este modelo cada sujeto puede tener una tasa de crecimiento diferente en la distancia dental a lo largo del tiempo. Observando el modelo, vemos como ahora 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 pendiente entre sujetos de **0.051**; obteniendo una mejora significativa. Este tipo de modelos es más realista cuando hay variabilidad individual en la evolución de la variable respuesta.

Este modelo es más flexible, ya que permite que tanto la intersección como la pendiente de la relación entre edad y distancia varíen entre individuos. Este último modelo generaliza la idea

que vimos en el capítulo anterior, donde ajustábamos una regresión por individuo (Figura 4). En aquel 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 eficiente, combinando la información de todos los individuos para obtener estimaciones más robustas, sin necesidad de 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 solo **efectos aleatorios** permite que cada sujeto tenga su propio intercepto, pero mantiene una pendiente común para todos. El **modelo mixto (LMM)** es el más completo, permitiendo que tanto la intersección 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.

3.2 Modelos Lineales Mixtos (LMM)

Son métodos y modelos estadísticos que sirven para analizar datos longitudinales cuando la variable respuesta sigue una distribución normal. Uno de sus aspectos más característicos lo indica Francisco Hernández-Barrera en su libro *Modelos mixtos con R* (Hernández-Barrera 2024), ya que se asume que existe una relación entre el vector de observaciones y las covariables. Se considera la técnica más eficaz cuando se trabaja con distribuciones normales en este campo ya que permite introducir efectos aleatorios y concretar la estructura de las correlaciones de los residuos del mismo sujeto; además de que puede emplearse con datos faltantes. Estos modelos nos permiten modelar la correlación entre observaciones dentro de una misma unidad e incluir covariables tanto a nivel individual como grupal. Los LMM permiten realizar una estimación precisa de la incertidumbre, respetando la dependencia entre observaciones. Por otro lado, su capacidad de generalización a estructuras de datos complejas es otro de los motivos por los cuales se recomienda su uso con datos longitudinales. Otra de sus ventajas es su flexibilidad para incluir efectos específicos por individuo o grupo; algo que veremos más adelante.

La ecuación para este tipo de modelos, en los que y_{ij} representa el momento j-ésimo del individuo i:

$$y_{ij} = \beta_{0i} + \sum_{k=1}^{K} \beta_{ki} x_{ijk} + e_{ij}$$

- x_{ijk} es el valor de la k-ésima variable independiente por parte del individuo i en la observación j.
- β_{0i} sigue $N(\beta_0, \sigma^2_{\beta_0})$; es el intercepto del modelo, que suele tener cierta varianza centrada en μ porque se supone aleatoria.

- β_{ki} sigue $N(\beta_k, \sigma^2_{\beta_k})$; son las pendientes o coeficientes de las variables independientes del modelo, que suelen ser aleatorias.

Los **efectos aleatorios** se representan mediante el vector formado por la constante y los coeficientes aleatorios del modelo. Nos permiten capturar la variabilidad entre individuos, y se escriben de esta forma:

$$\vec{\beta}_i = (\beta_{0i}, \beta_{1i}, ..., \beta_{Ki})^t \sim N(\vec{\beta}, \Omega)$$

Cabe destacar que los errores de un individuo, al no tener todos el mismo número de observaciones, son **independientes** de los efectos aleatorios.

Para ajustar un modelo lineal mixto, se tienen que disponer los datos de forma vertical. Una de las ventajas del LMM es su flexibilidad ya que no sólo permite especificar efectos aleatorios para evaluar la **variabilidad** de algunas variables entre los individuos, sino que también permite evaluar la **correlación** entre distintos datos longitudinales del mismo individuo. La constante y los coeficientes aleatorios tienen **homocedasticidad**, ya que la esperanza y la matriz de covarianzas es la misma para todos los individuos. Una de las características de los LMM es que introducen el concepto de **efectos fijos**, los cuales son la esperanza de los efectos aleatorios. Según Julian Faraway en *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* (Faraway 2006), un efecto fijo es una constante desconocida que intentaremos estimar a partir de los datos, mientras que los efectos aleatorios son variables aleatorias. De hecho, cuando un coeficiente no es aleatorio, se puede asumir que sigue una distribución normal con varianza cero; denominándolo fijo. En estos modelos, el número de efectos aleatorios es limitado, ya que no pueden superar en ningún caso el número de medidas que tenemos por individuo. Estos efectos inducen una **correlación** entre datos para el mismo individuo, pero dependiendo de su estructura la correlación sólo se puede obtener a partir de la correlación entre residuos; ya que si consideramos los coeficientes de variables **cambiantes** en el tiempo como aleatorios la correlación es distinta según los tiempos de las medidas, mientras que si consideramos los coeficientes de variables **constantes** en el tiempo inducimos **heterocedasticidad** entre individuos.

Una vez formulado el modelo, necesitamos estimar los parámetros, contando con los dos principales métodos de estimación en modelos mixtos. El primero es el método de Máxima verosimilitud (ML), que utiliza la función de verosimilitud completa del modelo. Estima tanto los efectos fijos como las varianzas de los efectos aleatorios. 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 el sesgo en la estimación de la varianza. Es el método preferido para comparar modelos con la misma estructura de efectos fijos pero distinta estructura de efectos aleatorios.

A la hora de trabajar con Modelos Lineales Mixtos, se puede trabajar de diferentes formas. Podemos establecer un modelo con la constante aleatoria y varios coeficientes fijos en el tiempo, en cuyo caso, si asumimos que los errores son independientes, tendríamos una correlación **constante** entre las variables del mismo individuo que no depende de la distancia entre las

medidas; lo que se denomina como coeficiente de correlación intraclase (ICC). Otra forma de definir estos modelos podría ser con la constante y los coeficientes aleatorios, donde, asumiendo independencia entre residuos, la correlación entre observaciones pasa a depender tanto del tiempo como de la distancia entre ellas. Sin embargo, pese a ser las dos buenas opciones, es preferible trabajar de otra forma para LMM.

Para empezar, no asumiremos independencia de los residuos; sino que trabajaremos con un modelo más general en el que contemos con el mayor número posible de efectos aleatorios correlacionados y fijos. A continuación, procederemos a simplificar el modelo a través de la significación de **efectos aleatorios**:

$$\begin{cases} H_0 : \sigma^2_{\beta_0} = 0 \\ H_1 : \sigma^2_{\beta_0} > 0 \end{cases}$$

Para comprobar que hay más de un efecto aleatorio significativo, se utilizan diferentes técnicas estadísticas para contrastar que se permite asumir que podemos rechazar la hipótesis nula: los efectos aleatorios tienen varianza igual a cero. En caso afirmativo, tenemos que contrastar que si su correlación es distinta de 0; para lo que tendremos que elegir la matriz de covarianzas de los efectos aleatorios:

$$\begin{cases} H_0 : \Omega = \begin{pmatrix} \sigma^2_{\beta_0} & 0 \\ 0 & \sigma^2_{\beta_1} \end{pmatrix} \\ H_1 : \Omega = \begin{pmatrix} \sigma^2_{\beta_0} & \sigma_{\beta_0\beta_1} \\ \sigma_{\beta_0\beta_1} & \sigma^2_{\beta_1} \end{pmatrix} \end{cases}$$

En este caso, utilizaremos un test de razón de verosimilitudes para escoger la estructura de covarianzas de los efectos aleatorios y de sus errores. Para ello, hay que tener en cuenta que los modelos estén **anidados**, es decir, que la matriz de covarianzas de los residuos de un modelo se expresen como un caso particular de la matriz de covarianzas de los residuos del otro modelo.

Una vez hemos terminado con los efectos aleatorios, procedemos a determinar la significación de los efectos fijos a través de dos métodos. Si queremos testear un sólo parámetro, utilizaremos el test de Wald en el que:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

En caso de querer testear más de un parámetro, utilizaremos un test de razón de verosimilitudes:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = 0 \\ H_1 : \text{alguno diferente de } 0 \end{cases}$$

Una vez hemos definido ya nuestro modelo, tenemos que realizar su validación a través de comprobar que se cumplen las asunciones sobre los residuos; al igual que hacíamos con Regresión Lineal Simple. Para poder asumir que el modelo es correcto, en el gráfico **residuos estandarizados vs valores predichos**, debería de aparecer una especie de nube de puntos en los que no haya ningún patrón ni ninguna tendencia aparente; mientras que en el **QQ-plot**,

si los residuos se encuentran alrededor de la diagonal sin seguir tampoco ningún patrón, podremos asumir que los residuos tienen normalidad. Para validar lod efectos aleatorios, podemos utilizar **Empirical Bayes Estimates** en lugar de asumir su normalidad.

3.3 Modelos Lineales Generalizados (GLM)

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 como la media de la variable respuesta y la combinación lineal de variables explicativas están relacionadas. Según Paul Roback y Julie Regler en el libro *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R*(Roback and Legler 2021), los GLM son una clase de modelos más amplia que tienen formas parecidas para sus varianzas, verosimilitudes y MLEs; generalizando la regresión lineal múltiple.

La **familia exponencial** suele tener 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 (location); mientras que ϕ es el **parámetro de dispersión** y representa la escala (scale). 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.

Para determinar si un modelo está basado en un único parámetro θ, tenemos que poder escribir su función de probabilidad de la siguiente forma:

$$f(y; \theta) = e^{[a(y)b(\theta)+c(\theta)+d(y)]}$$

Si el conjunto de posibles valores de entrada no depende de θ, la familia exponencial será de un parámetro. 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 fnción 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, podemos demostrar 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:

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

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, somos capaces de 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.

3.4 Modelos Lineales Generalizados Mixtos (GLMM)

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, extendemos 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, Gamma…), y la correlación entre observaciones repetidas para el mismo individuo, mediante efectos aleatorios.

La ecuación general de un GLMM es:

$$g\left(\mathbb{E}(y_{ij} \mid b_i)\right) = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ijk} + Z_{ij} b_i$$

Donde:

- y_{ij} es la respuesta del individuo i en la ocasión j.
- x_{ijk} es el valor de la variable explicativa k para ese individuo en ese momento.
- β_{0i} es el intercepto aleatorio.
- β_{ki} puede incluir efectos fijos o aleatorios.
- $g(\cdot)$ es la función de enlace, que conecta el valor esperado con la combinación lineal de predictores.

Dependiendo de la naturaleza de y_{ij}, usaremos distintos enlaces. Para datos binarios: enlace logit y distribución binomial; para datos de conteo: enlace log y distribución de Poisson; y para tiempos o proporciones: 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`.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: aprobado ~ horas + grupo + (1 | estudiante)
   Data: df_glmm

     AIC      BIC   logLik deviance df.resid
   529.0    545.8   -260.5    521.0      496

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.6296 -0.6712  0.3084  0.5411  2.1805

Random effects:
 Groups     Name        Variance Std.Dev.
 estudiante (Intercept) 0.8974   0.9473
Number of obs: 500, groups:  estudiante, 100

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.93155    0.27963  -3.331 0.000864 ***
horas        0.39577    0.05069   7.808  5.8e-15 ***
grupoB       0.26530    0.24055   1.103 0.270081
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
       (Intr) horas
horas  -0.710
grupoB -0.479  0.066
```

En este modelo GLMM (glmer) 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), con A como categoría de referencia). 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.

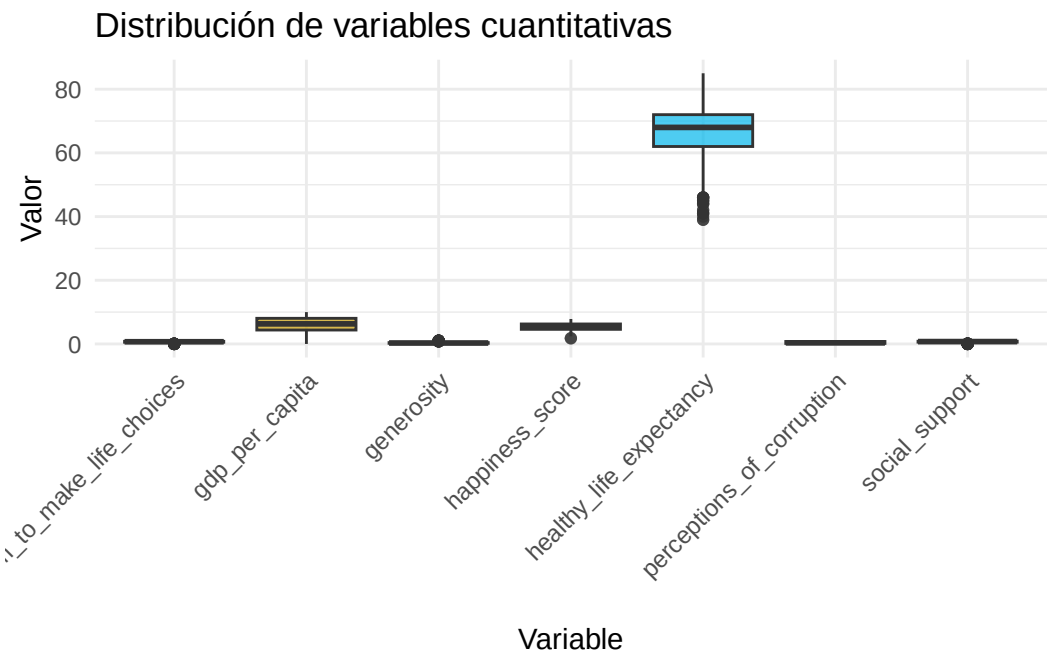
El intercepto aleatorio por estudiante tiene una desviación estándar de 0.9473, lo que indica que hay una variación importante en la propensión a aprobar entre estudiantes, incluso tras controlar por las horas de estudio y el grupo; lo que justifica el uso de un GLMM en lugar de un GLM clásico. Este modelo demuestra cómo un GLMM puede capturar variabilidad individual (entre estudiantes) y a la vez evaluar el efecto de factores fijos. El uso del modelo mixto es crucial: si ignoramos el efecto aleatorio por estudiante, estaríamos asumiendo que todos los estudiantes tienen la misma propensión a aprobar, lo cual claramente 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 jerárquicos, y permiten hacer inferencia poblacional y considerar la dependencia temporal. Por tanto, los GLMM constituyen una herramienta 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.

4 Análisis de la base de datos

El conjunto de datos World Happiness 2015-2024 recopila información sobre la felicidad percibida en diferentes países a lo largo de los años. Este dataset 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 nivel mundial. El dataset tiene una buena cobertura temporal, ya que abarca datos de 2015 a 2024 y permite analizar tendencias a lo largo del tiempo, pero también geográfica, dado que icluye información de diferentes países y regiones del mundo. Este dataset es ampliamente utilizado en estudios de bienestar, calidad de vida y políticas públicas, y contiene métricas económicas y sociales que permiten un análisis estadístico y comparativo. Cada fila del dataset 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 (escala de 0 a 10).
- GDP per capita: Producto Interno Bruto per cápita ajustado por poder adquisitivo.
- Social Support: Medida de apoyo social basado 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, según encuestas de percepción.
- 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. ## Análisis exploratorio inicial



Podemos ver como en la base de datos no hay valores faltantes, y a través de este análisis, vemos como tampoco hay valores atípicos. No obstante, para complementar nuestro análisis de la base de datos World Happiness, hemos considerado 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 democracia, los derechos políticos y las libertades influyen en la percepción de felicidad de las sociedades.

La primera base de datos que hemos considerado es "Freedom in the World", un informe anual de la organización Freedom House, que evalúa el estado de las libertades políticas y civiles en el mundo. Cada país es clasificado en función de indicadores de democracia, libertades individuales y derechos políticos. 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 la política y vivir sin restricciones autoritarias. Incorporar estos datos nos permitirá ver si existe una correlación entre los niveles de libertad y la felicidad percibida en cada país. Como esta base 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, similar al regional_indicator del dataset original.
- 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.

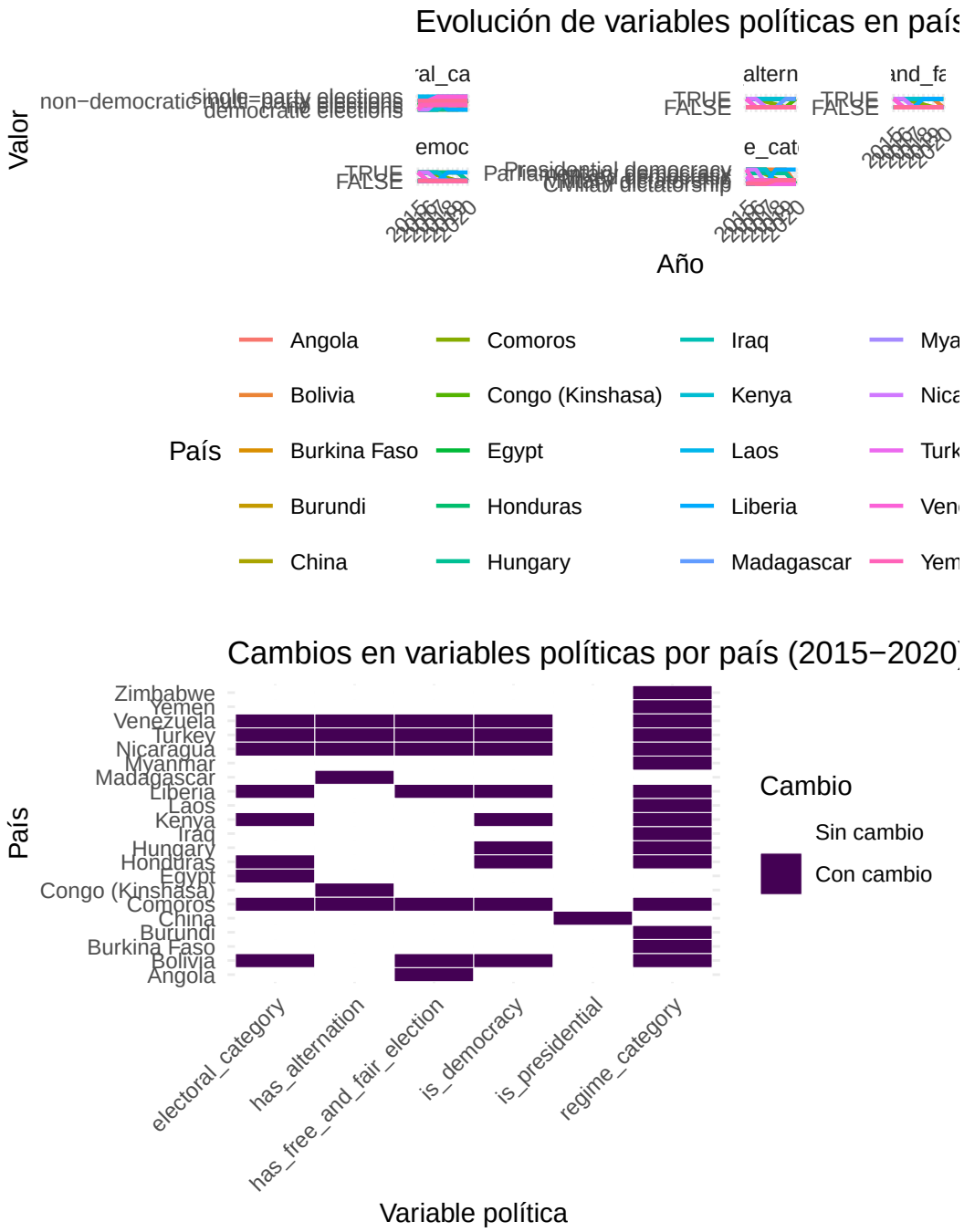
La otra base de datos que hemos elegido para nuestro análisis es "Democracy Data", un dataset que proviene del proyecto TidyTuesday y está basado en estudios académicos sobre democracia y regímenes políticos. Clasifica los países según su sistema de gobierno y proporciona información detallada sobre su historia política. Una de las características que tiene esta base de datos es que tiene datos hasta 2020, por lo que tenemos que considerar que, si vamos a trabajar con ella, tendremos las características completas de las observaciones en un período reducido. Dado que la felicidad no solo depende de factores económicos, sino también de la estabilidad política y la gobernanza, 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 pensamos 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 son libres y justas.
- has_alternation: Indica si existe alternancia en el poder.

Ambos datasets complementarán la base de datos de felicidad al agregar información sobre libertades civiles, derechos políticos y calidad democrática. Esto nos permitirá realizar un análisis más profundo y responder preguntas sobre la relación entre política, democracia y bienestar subjetivo en distintos países.

Una de las preguntas más interesantes al trabajar con datos longitudinales sobre países es si sus características políticas se mantienen estables o si sufren cambios significativos a lo largo de los años. Aunque la mayoría de los países presentan estabilidad en estas variables durante

el período 2015-2020, hemos identificado cambios en una serie de países que han experimentado transformaciones en su tipo de régimen, forma de gobierno o estatus democrático. A continuación, exploramos gráficamente cómo han evolucionado estos cambios en el tiempo.



El gráfico muestra de forma clara y visual la evolución temporal de varias variables políticas clave en los países que han experimentado cambios entre 2015 y 2020.

Si observamos los cambios en el tipo de elecciones, podemos ver que Venezuela presentó hasta 3 categorías distintas: pasó de elecciones democráticas a elecciones no democráticas multi-partido y luego a no tener elecciones directamente, reflejando el colapso progresivo del sistema electoral bajo el régimen de Maduro. Otros países como Honduras o Comoras alternan entre elecciones democráticas y elecciones no democráticas multi-partido, reflejando cuestionamientos sobre la transparencia y legitimidad de sus procesos electorales y evidenciando las dificultades para establecer un sistema electoral pluralista. En el caso de Honduras a partir de 2018, lo que coincide con la reelección altamente cuestionada de Juan Orlando Hernández, rodeada de denuncias de fraude y represión; mientras que Comoras en 2019, tras una serie de reformas constitucionales que favorecieron la consolidación del poder del presidente Azali Assoumani.

En cuanto a los cambios en la alternancia en el poder, vemos cómo en Nicaragua y Venezuela esta variable refleja la falta de alternancia, mostrando un patrón de concentración del poder ya que en años anteriores había alternancia, pero se pierde conforme los gobiernos se consolidan en el poder.

Si observamos los cambios en la existencia de elecciones libres y justas, vemos que Angola y Liberia reflejan mejoras en sus procesos electorales con la celebración de elecciones que fueron más competitivas y abiertas que en años anteriores, pero hay otros casos como Bolivia, que presenta un cambio negativo en 2019, año marcado por acusaciones de fraude electoral que desembocaron en la crisis política. Otro ejemplo que se puede destacar es Turquía, que deja de tener elecciones libres y justas en los últimos años, en línea con el creciente autoritarismo del presidente Erdoğan.

Observando la condición de democracia, vemos que Nicaragua dejó de ser considerada una democracia a partir de 2017, cuando se intensificó el control autoritario de Daniel Ortega sobre las instituciones y se reprimieron masivamente las protestas. Hungría también dejó de ser considerada democracia a partir de 2019, en consonancia con el cambio de régimen antes mencionado. Otro cambio lo observamos en Bolivia, que fue considerada democracia durante gran parte del período, pero dejó de serlo temporalmente tras la crisis política y la salida de Evo Morales en 2019.

Si observamos la forma de gobierno presidencial, vemos que el único Estado que cambia es China: aunque siempre fue una dictadura de partido único, el dataset refleja un cambio en esta variable posiblemente por una reinterpretación metodológica, ya que a partir de 2019 es considerada presidencial. Es destacable que esta variable apenas presenta cambios (salvo China), lo cual sugiere que la forma de gobierno en términos institucionales es más estable que otras dimensiones como la calidad electoral o la democracia sustantiva. Esto puede ayudar a entender que muchos procesos de regresión democrática no implican reformas estructurales del sistema político, sino más bien una transformación interna de las reglas del juego bajo el mismo marco institucional.

Por último, si observamos los cambios en el tipo de régimen, vemos como Hungría pasó de ser una democracia parlamentaria a una dictadura civil, lo cual refleja el creciente autoritarismo del gobierno de Viktor Orbán y el deterioro de la separación de poderes y la libertad de prensa. Sin embargo, hay casos como Liberia que sucede todo lo contrario: en los primeros años del periodo considerado, Liberia fue clasificada como una dictadura civil, reflejo del mandato de Ellen Johnson Sirleaf, que, aunque existía cierta estructura institucional, no se garantizaban plenamente principios básicos de la democracia como elecciones libres o alternancia real de poder. Sin embargo, a partir de 2018, Liberia pasa a ser clasificada como una democracia presidencial,cambio que coincide con la llegada al poder del exfutbolista George Weah, quien fue elegido democráticamente en un proceso que representó la primera transición pacífica de poder entre dos presidentes electos en el país desde 1944. Esta transición marcó un punto de inflexión político, reflejando una mejora significativa en las instituciones democráticas del país. En Burundi, el cambio en el tipo de régimen refleja el deterioro institucional que siguió a la decisión de Pierre Nkurunziza de presentarse a un tercer mandato en 2015, desatando una grave crisis política. El país pasa a ser clasificado como dictadura civil, reflejando la suspensión de libertades y el cierre del espacio cívico. Este cambio es particularmente relevante porque anticipa un patrón que luego se repite en otros países: el uso del poder electoral como instrumento de legitimación de autoritarismo. Otro caso a destacar es Zimbabwe, ya que es interesante porque, aunque solo presenta un cambio en regime_category, este se produce tras la caída de Robert Mugabe en 2017, una transición que fue vista por algunos sectores como una oportunidad para abrir un nuevo ciclo democrático. Sin embargo, el hecho de que no haya cambios en otras variables como is_democracy o has_free_and_fair_election refleja que el cambio de liderazgo no implicó necesariamente una mejora sustantiva en la calidad del régimen.

Además de los ejemplos más visibles, hay otros países que presentan cambios más sutiles pero igualmente significativos. Por ejemplo, Madagascar muestra una alteración en la variable has_alternation, indicando un momento de alternancia que podría asociarse con las elecciones de 2018, cuando Andry Rajoelina volvió al poder tras haberlo ocupado anteriormente como presidente de transición. Este cambio sugiere un contexto político volátil en el que las alternancias no siempre reflejan procesos plenamente democráticos, sino que pueden estar vinculadas a luchas internas o acuerdos de élites.

Otro caso particular es Comoras, que presenta múltiples cambios en variables como el tipo de elecciones, la democracia y el tipo de régimen. Esto refleja un proceso de consolidación autoritaria que ha sido documentado por organismos internacionales tras la reforma constitucional de 2018 y las elecciones de 2019, en las que se concentró el poder presidencial y se limitó la oposición. Comoras es un ejemplo claro de regresión democrática multidimensional, donde no solo se pierde calidad electoral, sino que se transforman también las estructuras institucionales.

En Kenya, se observan cambios tanto en is_democracy como en electoral_category y regime_category, lo cual es coherente con un contexto de avances y retrocesos en la calidad democrática. A pesar de tener elecciones relativamente regulares, las denuncias de fraude,

violencia post-electoral y polarización han afectado la credibilidad del sistema. El dataset parece capturar bien esa inestabilidad institucional, mostrando a Kenya como un país en constante disputa entre aperturas democráticas y tendencias autoritarias.

Este gráfico refuerza la idea de que varios países han experimentado retrocesos democráticos significativos, especialmente en torno a elecciones libres, alternancia y la clasificación del régimen. Además, evidencia que estos cambios no son simultáneos: mientras algunos países cambian en 2016, otros lo hacen en 2018 o 2020, lo cual permite contextualizar los cambios políticos con eventos históricos concretos en cada nación. Este tipo de análisis temporal no solo es útil para identificar patrones políticos, sino también para cruzar estos cambios con la percepción de felicidad de la población y evaluar si existe alguna asociación relevante entre ambas dimensiones.

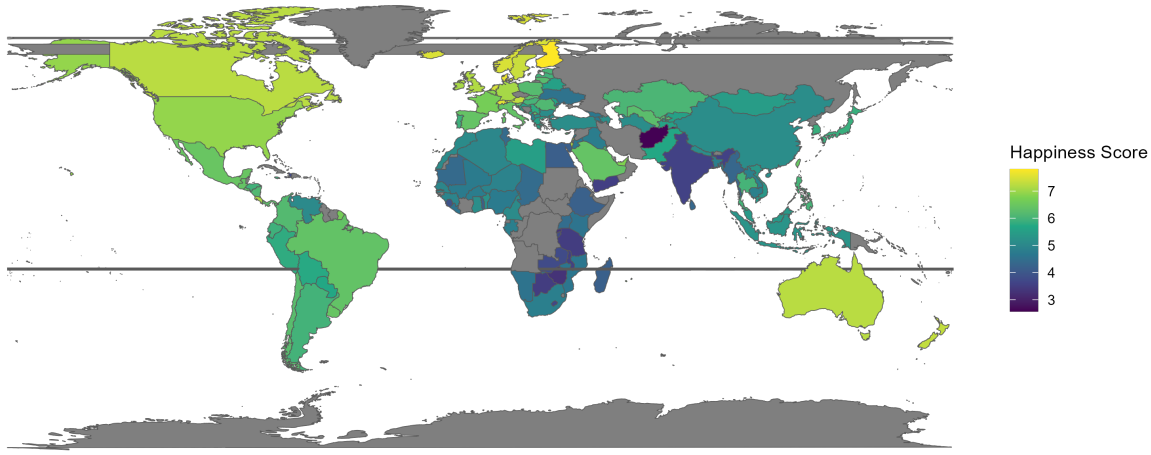
Dado que uno de los objetivos principales de este análisis es estudiar la evolución de la felicidad y sus determinantes a lo largo del tiempo, necesitamos trabajar con una base de datos que tenga cobertura completa para el periodo 2015–2024. En este sentido, hemos optado por utilizar la base de datos formada por los datasets "World Happiness" y "Freedom in the World", excluyendo la base de datos de "Democracy Data".

Esta decisión se debe a que la base de datos de democracia únicamente contiene información hasta el año 2020, lo cual limitaría drásticamente el análisis longitudinal si trabajásemos únicamente con ella. Por lo tanto, para poder aplicar técnicas de medidas repetidas, donde analizamos cómo varían las observaciones de un mismo país a lo largo de varios años, necesitamos una base con continuidad temporal.

No obstante, la base de "Democracy Data" no se descarta completamente. Se utilizará como complemento para proporcionar una fotografía institucional y política de los países, especialmente útil para interpretar fenómenos relevantes detectados durante el análisis longitudinal. Esta fotografía será especialmente relevante en los años en los que se observe una caída o subida abrupta en la felicidad, y nos permitirá contextualizar esos movimientos a través del tipo de régimen, la existencia de elecciones libres o la alternancia en el poder, entre otros factores.

Ahora que ya tenemos preparado el dataset, es momento de empezar a ver la evolución de la variable objetivo a lo largo del tiempo. Para ello, podemos hacer uso de mapas que nos ayudan a interpretar la información.

Mapa de Felicidad por País (Sin democracia, 2020)



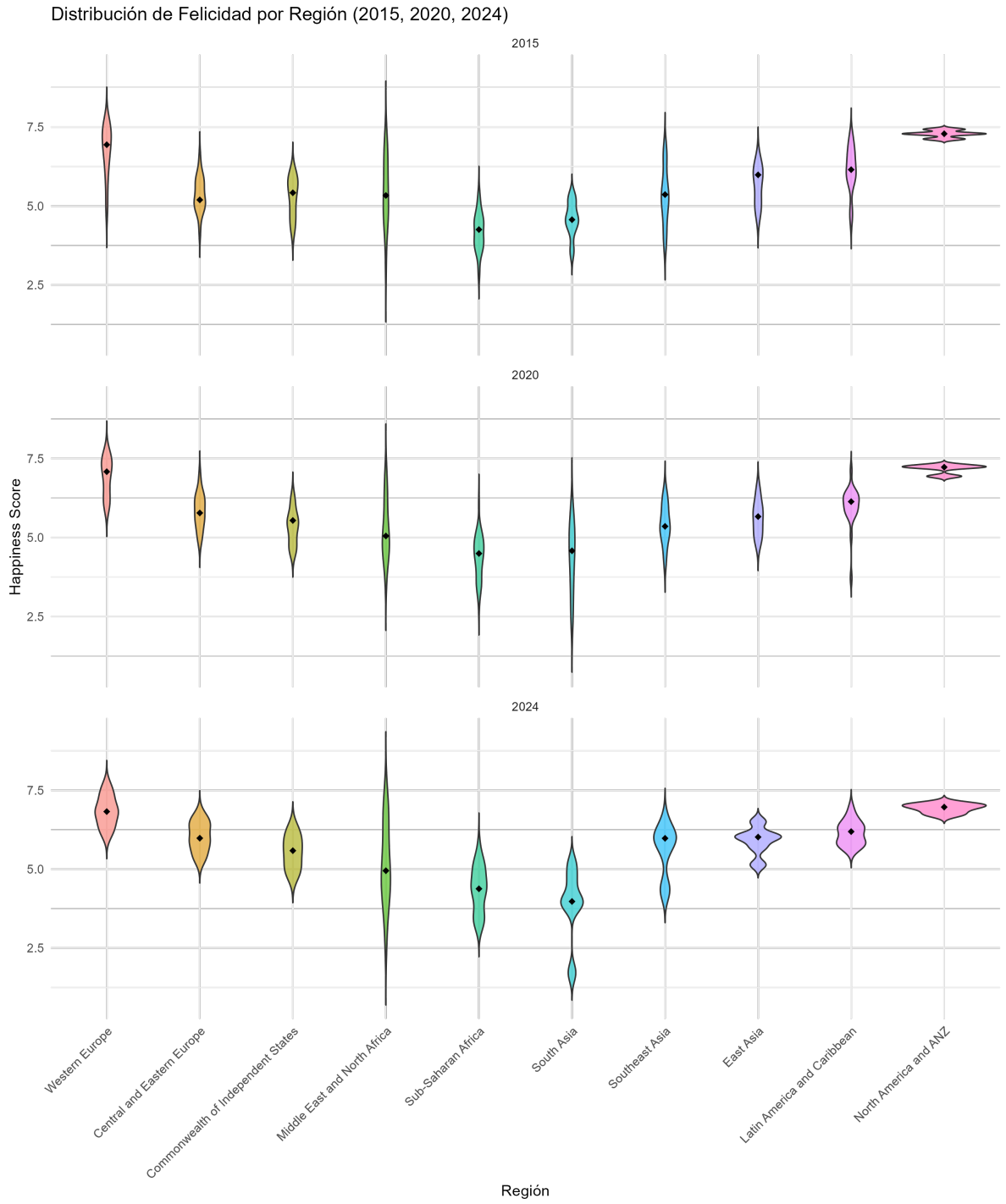
El mapa de felicidad por país en 2020 muestra una clara heterogeneidad geográfica en la percepción del bienestar, capturada a través del Happiness Score. Se observan regiones del norte global, como Europa Occidental, América del Norte y Oceanía, con los niveles más altos de felicidad (colores amarillos y verde claro), mientras que regiones del sur global, 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, Noruega y Nueva Zelanda destacan con las puntuaciones más altas, reflejando contextos estables y altos niveles de desarrollo económico y social. En contraste, países como Zimbabwe, Sudán del Sur o Afganistán presentan los niveles más bajos, lo cual es coherente con contextos de conflicto, pobreza o inestabilidad política.

Este mapa no solo permite identificar diferencias entre regiones, sino también subrayar patrones estructurales: América Latina, por ejemplo, muestra un nivel medio de felicidad, con cierta variabilidad entre países. Es especialmente útil para detectar casos anómalos, como países con puntuaciones bajas en regiones generalmente altas o viceversa.

En conjunto, el gráfico evidencia de forma visual el impacto que pueden tener factores estructurales como el desarrollo, la gobernanza o la estabilidad en la percepción subjetiva del bienestar de los ciudadanos a nivel global.

Por otro lado, también podemos utilizar gráficos de violines. En este caso, podemos ver cómo la puntuación de la felicidad varía en función de la región.



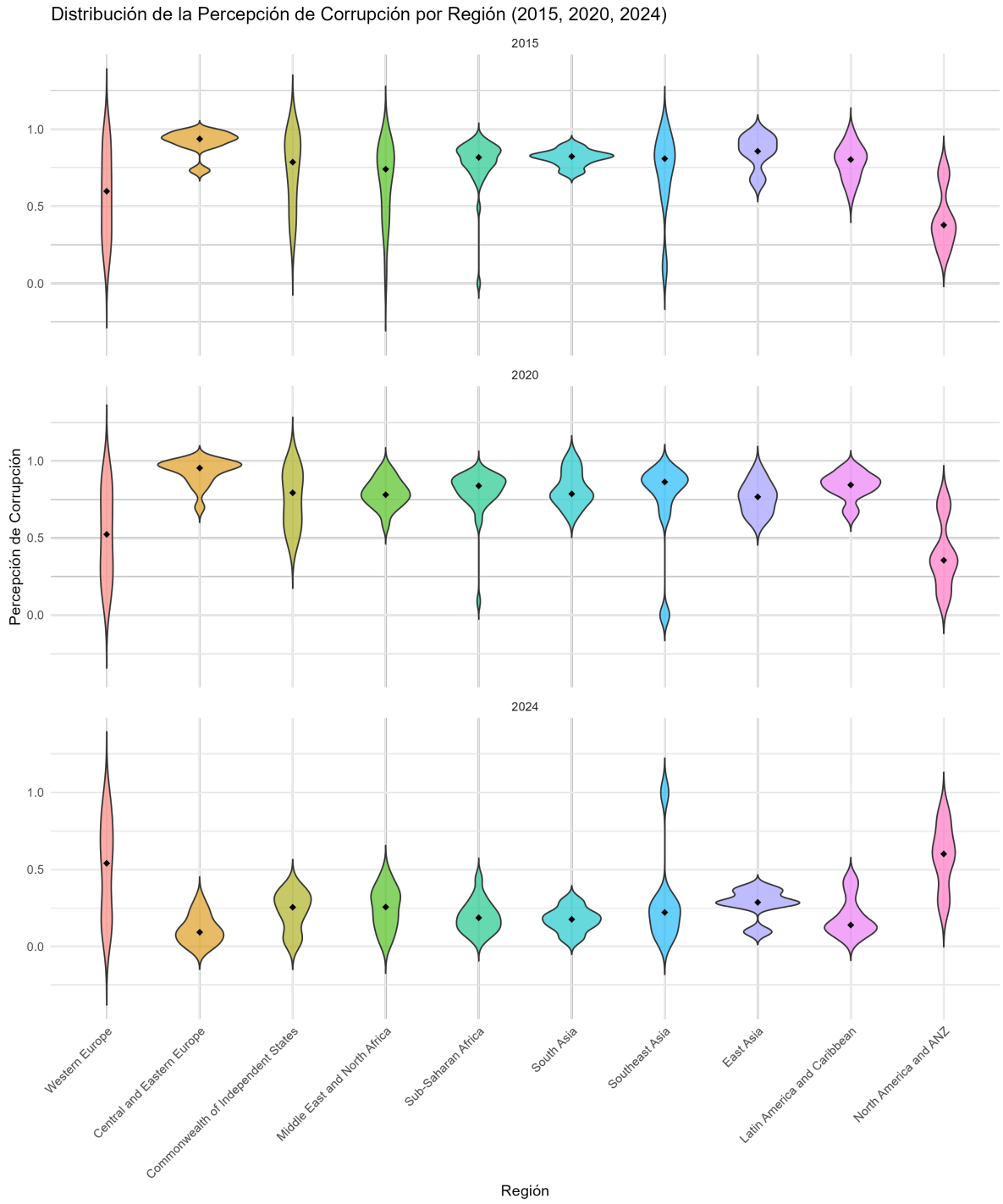
Si nos fijamos en la mediana, que se muestra con un rombo negro en cada violín, vemos que hay regiones como Europa Occidental o Norte América y Australia que mantienen medianas

altas y estables a lo largo del tiempo, mientras que el Sur de Asia y África Sub-Sahariana tienen medianas bajas, aunque relativamente estables.

En estos gráficos también se puede observar cierta bimodalidad, ya que algunas regiones, como Latino América y el Caribe o el Oriente Medio y África del Norte muestran dos modos (zonas más anchas del violín), lo que sugiere heterogeneidad interna: algunos países con altos niveles de felicidad conviven con otros con puntuaciones mucho más bajas. Asia del Sur presenta una distribución muy desigual en 2024, con un notable ensanchamiento en la parte inferior del gráfico, reflejando una baja felicidad en ciertos países (como Afganistán).

En general, la forma de las distribuciones se mantiene similar entre 2015 y 2024 en muchas regiones, pero algunas muestran mayor dispersión, sobre todo en el continente asiático, ya que Oriente Medio y Asia del Sur sufren una caída considerable en la felicidad. Regiones como Europa Occidental y Norte América y Australia presentan distribuciones compactas y felicidad más alta y estable, lo que refleja mayor consistencia en el bienestar subjetivo de sus poblaciones.

Otra de las variables que nos puede ayudar a interpretar la información es la percepción de la corrupción según las diferentes regiones.



El gráfico de violines permite observar la evolución de la percepción de corrupción en las distintas regiones a lo largo del tiempo, mostrando tanto la mediana como la distribución

completa de los datos. Lo más destacable es la aparición de una clara bimodalidad en algunas regiones a partir de 2020 y especialmente en 2024, lo que sugiere una divergencia creciente entre países dentro de la misma región.

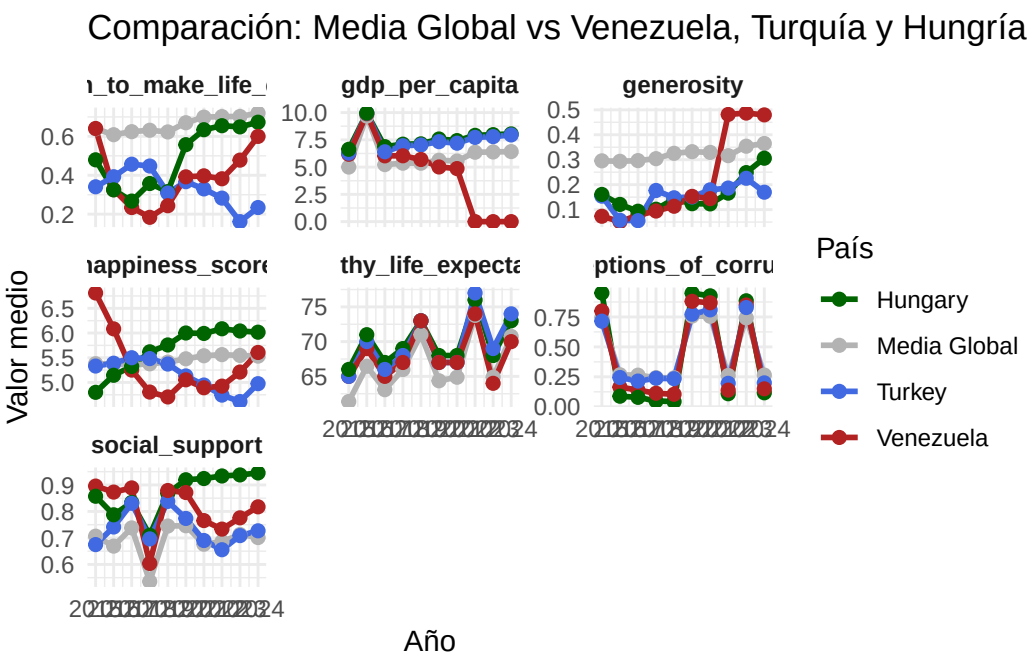
Europa Occidental mantiene una distribución relativamente estable a lo largo del tiempo, con una percepción de corrupción baja en general (valores altos), aunque en 2024 aparece una mayor dispersión que en 2015 y 2020. Sin embargo, en Europa Central y Oriental se observa un patrón inverso. En 2015 y 2020, la percepción es bastante alta (valores cercanos a 1), pero en 2024 se evidencia una caída brusca en la percepción positiva (valores muy bajos), lo que indica un empeoramiento de la percepción ciudadana sobre la corrupción en algunos países, como puede ser el caso de Hungría o Polonia.

Commonwealth of Independent States muestra una bimodalidad clara en 2024, mientras que en 2015 y 2020 predominaba un grupo más homogéneo. Esto sugiere que algunos países han experimentado mejoras o retrocesos significativos, diferenciándose del resto. América Latina y el Caribe también muestra un cambio claro: de una percepción moderadamente positiva en 2015 a una distribución más polarizada en 2020 y 2024, reflejando posiblemente tensiones políticas y escándalos de corrupción en países como Venezuela, Nicaragua o Bolivia.

En Asia del Este se observa una distribución muy compacta en 2015 y 2020, pero con una clara ruptura en 2024, lo que podría estar relacionado con dinámicas opuestas en países como China (control político férreo) frente a otros con gobiernos más abiertos. Por el contrario, regiones como Asia del Sur y África Sub-Sahariana muestran distribuciones bajas y más estables, aunque con una ligera mejora hacia 2024 en la percepción media.

Este análisis permite detectar no solo la evolución de la percepción media, sino también la heterogeneidad creciente dentro de regiones, y puede servir de base para cruzar estos patrones con los cambios políticos identificados previamente, evaluando si el deterioro institucional y la concentración del poder están efectivamente correlacionados con una mayor percepción de corrupción entre la ciudadanía.

Con el objetivo de comprender cómo ha cambiado la situación global en términos de bienestar subjetivo y factores asociados, analizamos a continuación la evolución promedio anual de las principales variables del informe de felicidad. Esta visualización nos permitirá detectar tendencias crecientes o decrecientes y posibles impactos globales, como crisis políticas, sanitarias o económicas.



La evolución anual del promedio de variables de interés refleja tendencias globales relevantes, pero su comparación con países concretos permite matizar dichas dinámicas y comprender mejor cómo contextos políticos específicos afectan al bienestar percibido por la población.

En primer lugar, Venezuela se desmarca de forma clara en varias variables. El PIB per cápita presenta una caída drástica desde 2017, situándose muy por debajo de la media global a partir de 2018, lo que refleja la aguda crisis económica que atraviesa el país. Esta caída va acompañada de un descenso en la libertad para tomar decisiones y en la felicidad percibida, lo que sugiere un deterioro generalizado del bienestar. Sin embargo, resulta llamativo el aumento pronunciado en la generosidad a partir de 2021, superando incluso a la media global. Este fenómeno podría estar vinculado a la solidaridad comunitaria surgida ante la crisis prolongada, así como a cambios metodológicos en la forma en la que se capta esta variable en contextos de alta inestabilidad.

Por otro lado, Turquía muestra una tendencia decreciente en la libertad para tomar decisiones, especialmente desde 2018, coincidiendo con el fortalecimiento del poder ejecutivo bajo el liderazgo de Recep Tayyip Erdoğan. A pesar de mantener un PIB per cápita relativamente estable y una esperanza de vida en línea con la media global, la percepción de corrupción se mantiene sistemáticamente alta, lo cual refuerza la idea de un deterioro institucional progresivo. La felicidad percibida en Turquía permanece por debajo de la media, lo que puede reflejar un desencanto social persistente.

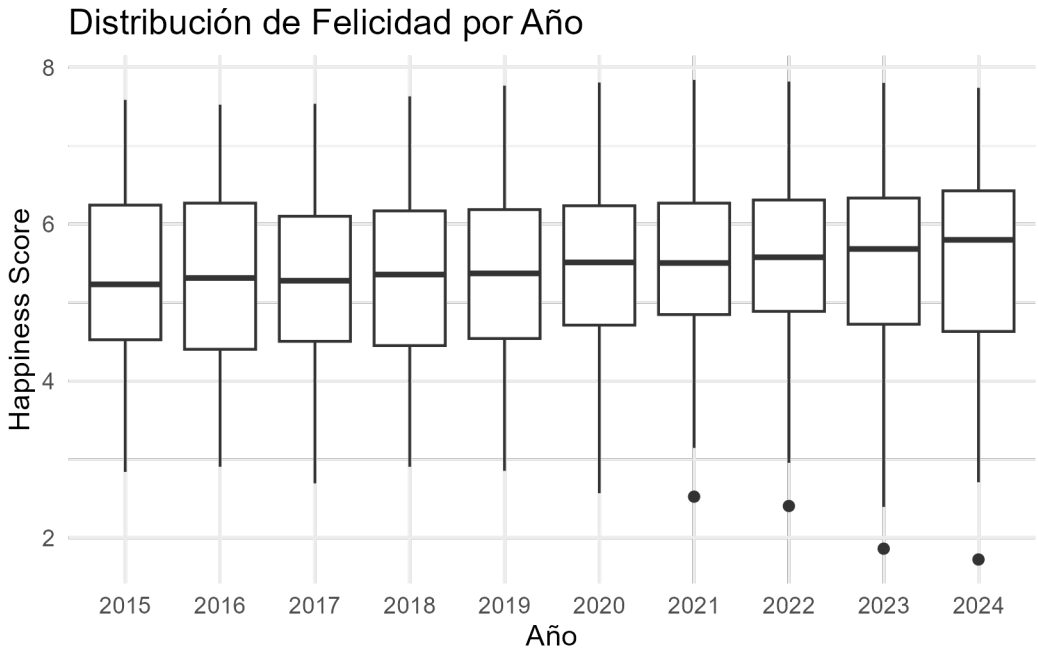
En el caso de Hungría, se observa una situación más ambivalente. Por un lado, variables como la esperanza de vida, el apoyo social y la felicidad percibida se mantienen por encima de la media global, lo que sugiere una cierta estabilidad material y comunitaria. No obstante, la

libertad para tomar decisiones experimenta un estancamiento, y la percepción de corrupción es elevada, lo cual coincide con el proceso de retroceso democrático documentado en el país desde mediados de la década de 2010. Este contraste pone de manifiesto cómo un país puede mantener ciertos niveles de bienestar mientras erosiona sus instituciones democráticas.

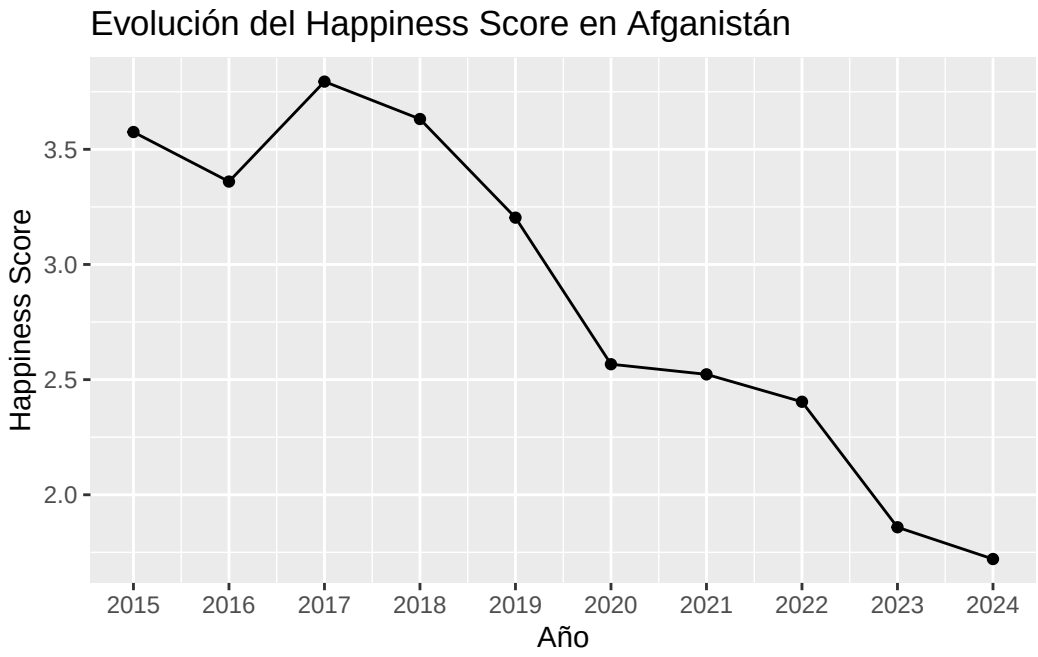
Este análisis complementa el estudio previo de la evolución global. Ya habíamos observado que, en conjunto, variables como la libertad para tomar decisiones y la generosidad presentan una evolución creciente, posiblemente impulsada por procesos de recuperación post-pandemia o cambios culturales. También destacaba un pico atípico en el PIB per cápita en 2016, probablemente debido a valores extremos, y una tendencia general de aumento en la felicidad percibida hasta 2022, seguida de un leve descenso. Por el contrario, la percepción de corrupción es más errática, con oscilaciones abruptas entre años, lo que refleja fuertes diferencias entre países.

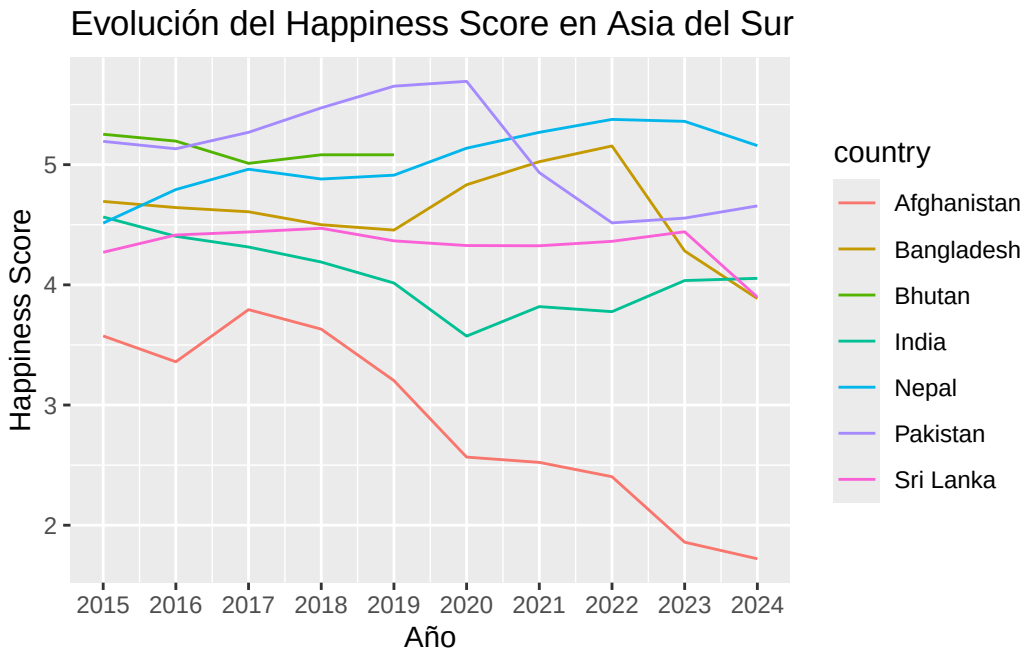
La comparación entre países permite, por tanto, observar cómo procesos políticos específicos (autoritarismo, crisis económica, reformas institucionales) se reflejan en los indicadores de bienestar. Venezuela, Turquía y Hungría muestran trayectorias divergentes respecto a la media global, lo cual subraya la necesidad de un análisis desagregado para captar la complejidad del bienestar en contextos políticamente inestables o en transformación.

Otra de las características que tenemos que analizar en esta base de datos es la existencia de valores atípicos que puedan afectar a nuestro análisis.



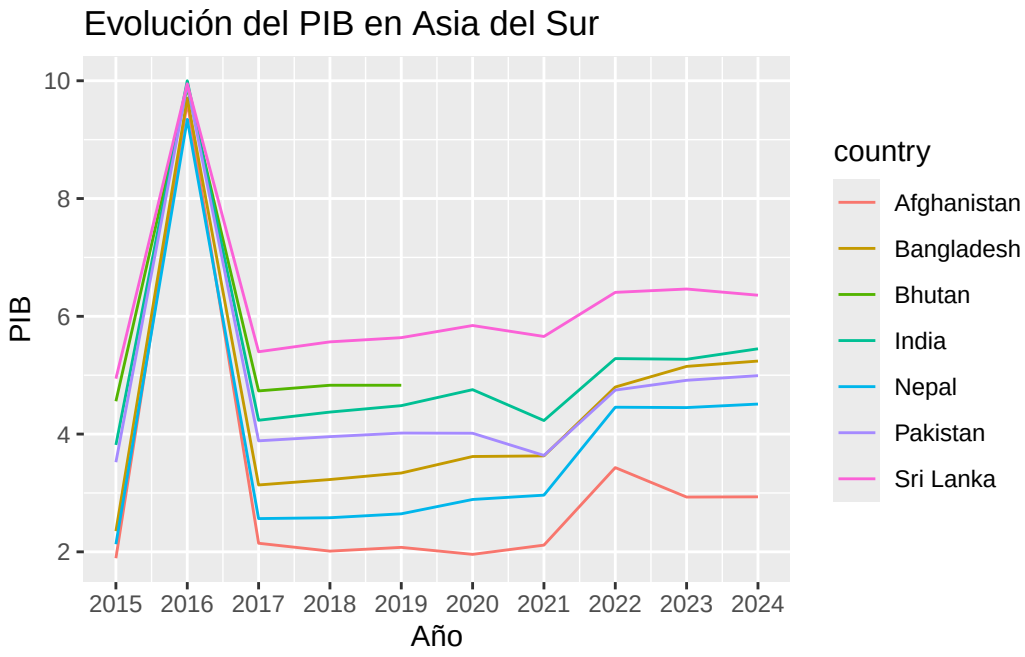
De 2015 a 2020 no vemos valores atípicos, pero en 2021 vemos un valor atípico de 2.52, en 2022 de 2.40, en 2023 de 1.86 y en 2024 de 1.72. Todos estos valores atípicos corresponden con Afganistán. Vamos a entrar en más detalle para ver cuál es la evolución de dicho país.

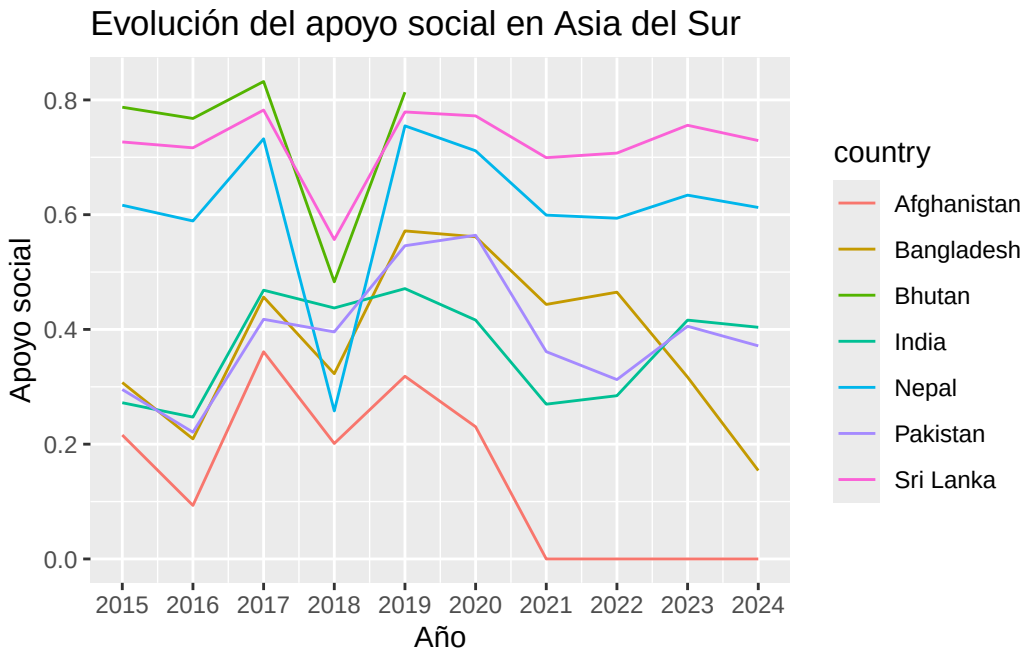


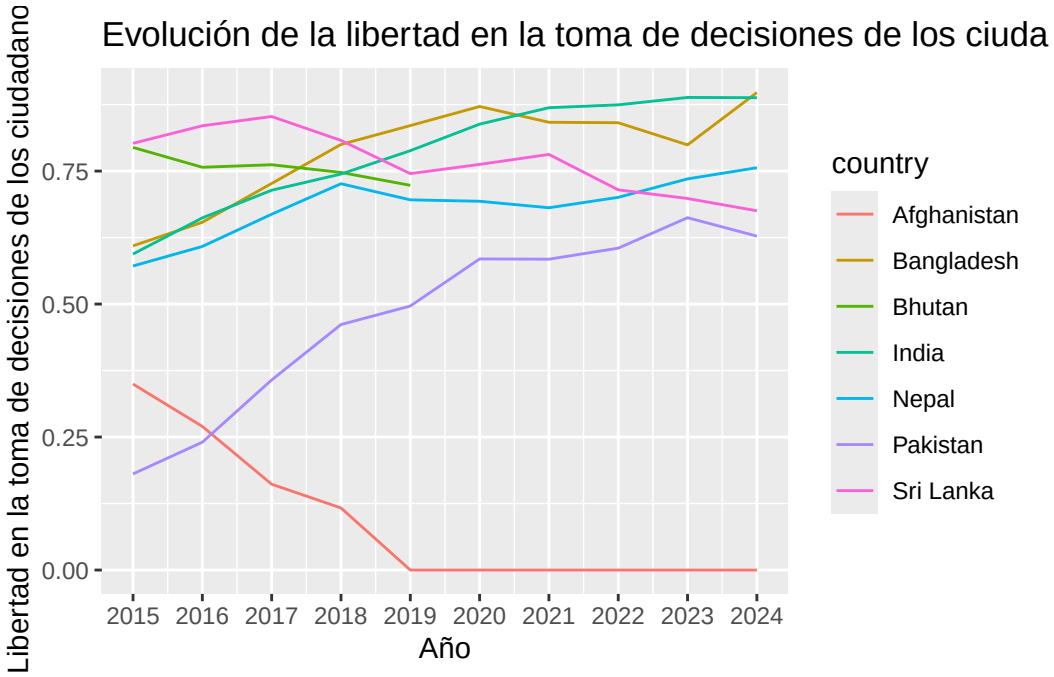


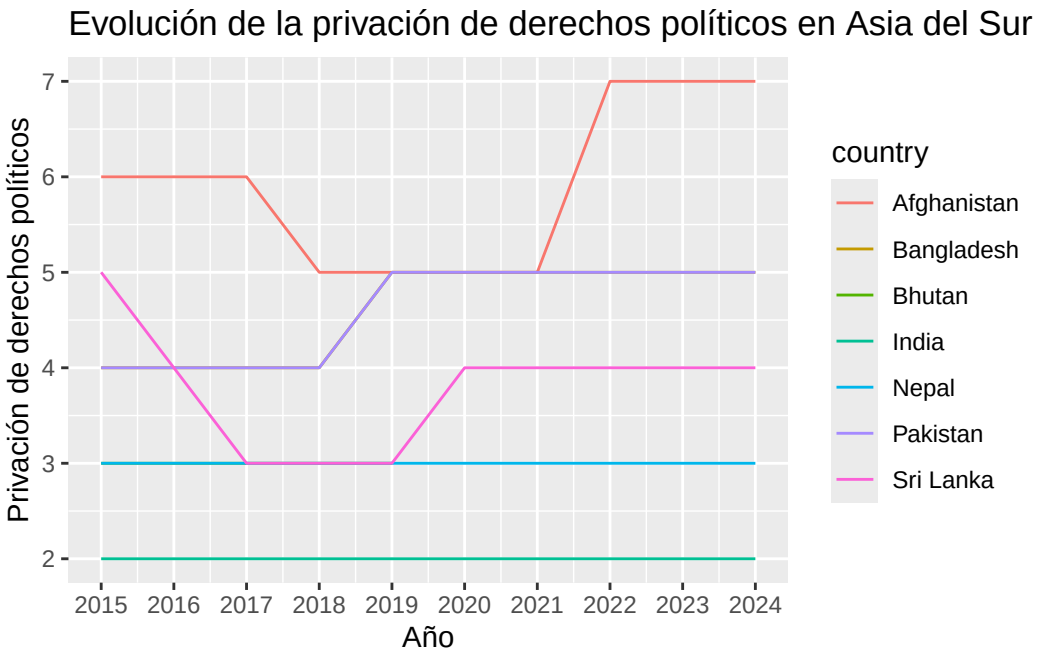
Afganistán sufre una gran caída de la felicidad a partir de 2017 y si lo comparamos con países de su región es el que menos puntuación tiene por bastante diferencia. Esto puede haberse dado por diversos factores, como el constante estado de guerra y conflicto en el que se ha encontrado el país, la presencia de los talibanes y otros grupos armados que han aumentado

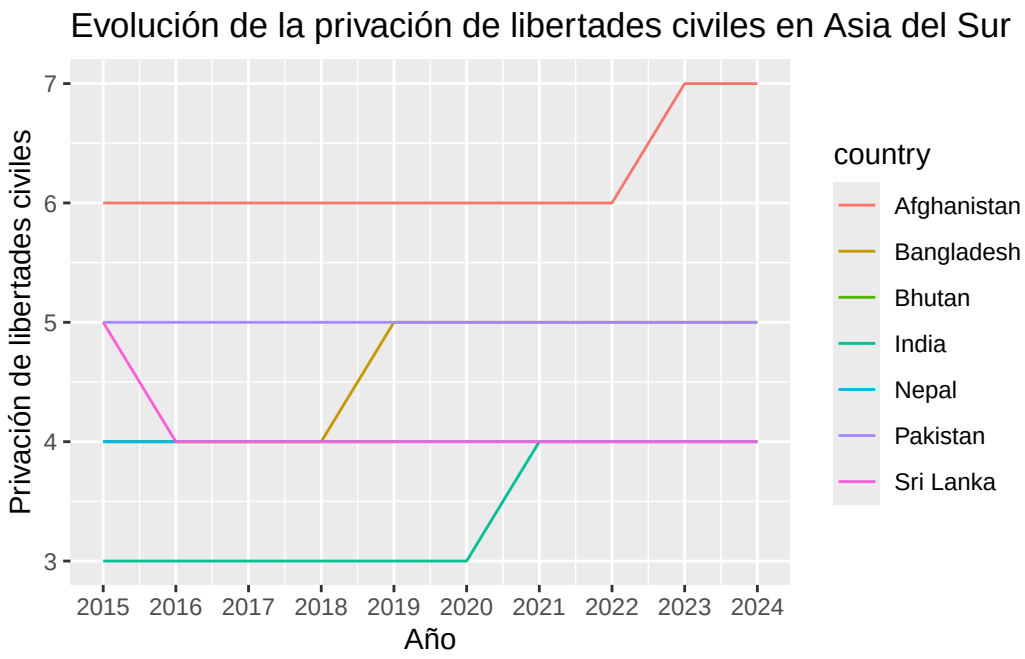
la violencia y el temor en la población, la enorme tasa de pobreza... Vamos a comparar la evolución de Afganistán con sus países vecinos para ver realmente qué puede estar afectando a tal bajada de la felicidad.






Evolución de la libertad en la toma de decisiones de los ciuda
Libertad en la toma de decisiones de los ciudadano
0.75
0.50
0.25
0.00
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Año
country
Afghanistan
Bangladesh
Bhutan
India
Nepal
Pakistan
Sri Lanka


Evolución de la privación de derechos políticos en Asia del Sur
Privación de derechos políticos
7
6
5
4
3
2
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Año
country
Afghanistan
Bangladesh
Bhutan
India
Nepal
Pakistan
Sri Lanka



Afganistán es el país con PIB más bajo, pero tampoco hay mucha diferencia con el resto; evolucionando acorde con sus países vecinos. Si nos fijamos en el apoyo social, aunque sigue la tendencia de sus vecinos, vemos cómo en el caso de Afganistán el apoyo social baja estrepitosamente hasta el punto de decir que es nulo a partir de 2021; justo cuando los talibanes toman el poder, lo cual ha podido afectar negativamente a la felicidad de los ciudadanos. Afganistán es, con gran diferencia, el país de Asia del Sur donde los ciudadanos tienen menos libertad para tomar decisiones; privándoles de una libertad que afecta de manera directa al descontento de los ciudadanos. Si nos fijamos en el resto de gráficas, vemos cómo Afganistán es el país donde más se ha privado de sus derechos políticos y donde menor libertad civil hay en todo Asia del Sur; lo cual explica por qué Afganistán tiene una puntuación de la felicidad tan baja en comparación con el resto del mundo.

4.1 Regresión Lineal Múltiple

Al principio, podemos observar que hay bastantes variables que pueden ser buenas predictoras del happiness_score. No obstante, en la selección de variables explicativas, vemos que el mejor modelo es aquel que cuenta con la percepción de la corrupción, la esperanza de vida, la libertad de tomar decisiones y el apoyo social. Al hacer el diagnóstico del modelo, vemos que el modelo no es fiable ya que los errores no tienen homocedasticidad ni normalidad.

Referencias

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

Hernández-Barrera, Francisco. 2024. "Modelos Mixtos Con r." 2024. https://fhernanb.github.io/libro_modelos_mixtos/.

Roback, Paul, and 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/.