UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Facultad de Administración y Dirección de Empresas /UPV

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Facultad de Administración y Dirección de Empresas

Modelos predictivos de niveles de estrés para mujeres en el climaterio a través de factores socio-económicos

Trabajo Fin de Grado

Grado en Administración y Dirección de Empresas

AUTOR/A: Borrás Asensio, Sofía

Tutor/a: Martínez Minaya, Joaquín

Cotutor/a: Vallada Regalado, Eva

CURSO ACADÉMICO: 2024/2025

A toda la gente que me ha apoyado durante este tiempo y sobre todo a mis tutores que no han dudado de mí en ningún momento, gracias por confiar.

RESUMEN

El presente Trabajo de Fin de Grado se enfoca en el análisis del nivel de cortisol en sangre, de mujeres residentes en la provincia de Valencia. Se examinarán diversas variables que podrían afectar a este nivel de cortisol como podrían ser la zona geográfica en la que residen, los niveles de estudio o el salario medio de las participantes. Para ello, en primer lugar se construirá una base de datos utilizando la información recopilada en la Fundació per al Foment de la Investigació Sanitària i Biomèdica de la Comunitat Valenciana (Fisabio). Una vez recopilados, se procederá al preprocesamiento y la transformación de los datos, con el objetivo de estandarizarlos en un formato compatible con R Studio. Posteriormente, se llevará a cabo un análisis exhaustivo de cada variable, evaluando su relación con la variable de interés, el nivel de cortisol en sangre. Tras comprender el panorama general, se desarrollará un modelo de regresión con el propósito de predecir los niveles de estrés de las mujeres a partir de un conjunto de variables predictoras. Inicialmente, se implementará un modelo de regresión múltiple sencillo, y se explorarán algoritmos más sofisticados con el objetivo de minimizar el margen de error en las predicciones.

RESUM

El present Treball de Fi de Grau s'enfoca en l'anàlisi del nivell de cortisol en sang, de dones residents a la província de València. S'examinaran diverses variables que podrien afectar este nivell de cortisol com podrien ser la zona geogràfica en la qual residixen, els nivells d'estudi o el salari mitjà de les participants. Per a això, en primer lloc es construirà una base de dades utilitzant la informació recopilada en la Fundació per al Foment de la Investigació Sanitària i Biomèdica de la Comunitat Valenciana (Fisabio). Una vegada recopilats, es procedirà al preprocessament i la transformació de les dades, amb l'objectiu d'estandarditzar-los en un format compatible amb R *Studio. Posteriorment, es durà a terme una anàlisi exhaustiva de cada variable, avaluant la seua relació amb la variable d'interés, el nivell de cortisol en sang. Després de comprendre el panorama general, es desenvoluparà un model de regressió amb el propòsit de predir els nivells d'estrés de les dones a partir d'un conjunt de variables predictores. Inicialment, s'implementarà un model de regressió múltiple senzill, i s'exploraran algorismes més sofisticats amb l'objectiu de minimitzar el marge d'error en les prediccions.

ABSTRACT

This Thesis focuses on the analysis of the level of cortisol in the blood of women living in the province of Valencia. Several variables that could affect this level of cortisol will be examined, such as the geographical area in which they live, the level of education or the average salary of the participants. To do this, a database will first be constructed using information collected at the Fundació per al Foment de la Investigació Sanitària i Biomèdica de la Comunitat Valenciana (Fisabio). Once collected, the data will be pre-processed and transformed in order to standardise them in a format compatible with R Studio. Subsequently, a comprehensive analysis of each variable will be carried out, assessing its relationship with the variable of interest, blood cortisol level. After understanding the big picture, a regression model will be developed for the purpose of predicting women's stress levels from a set of predictor variables. Initially, a simple multiple regression model will be implemented, and more sophisticated algorithms will be explored with the aim of minimising the margin of error in the predictions.

Tabla de contenidos

Índice de tablas

Índice de figuras

1. Introducción

1.1. Contexto y justificación

Actualmente, vivimos en una sociedad en la que los niveles de estrés son extremadamente elevados, afectando a un amplio sector de la población. Según la Organización Mundial de la Salud (OMS, 2023), el estrés se define como “el estado de preocupación o tensión mental generado por una situación difícil”. Estas situaciones, pueden ser derivadas de diversas razones, como preocupaciones económicas, estrés laboral, estilos de vida poco saludables, etc.

El presente Trabajo de Fin de Grado (TFG) tiene como objetivo principal la extracción y análisis de modelos predictivos de los niveles de estrés en las mujeres, teniendo en cuenta tanto sus características socio-económicas como los niveles de cortisol en sangre. Para ello, se integran datos de dos importantes proyectos: INMA - Infancia y Medio Ambiente (https://www.proyectoinma.org/), un estudio longitudinal de base poblacional que involucra a 239 mujeres de clase social media o media-alta, y PAPILONG, que incluye a 132 mujeres en riesgo de pobreza y exclusión social, usuarias de diversas ONG en Valencia. Ambos proyectos ofrecen una rica fuente de datos para este análisis, con reclutamientos realizados entre 2019 y 2022.

Es importante resaltar que el estrés no es un fenómeno aislado, sino que está influenciado por múltiples variables que pueden tener un impacto significativo en su manifestación. Algunas de las variables más importantes que se encuentran en las características socio-económicas son el nivel de ingresos, el acceso a la educación, las condiciones laborales y los hábitos que tenga el individuo. Estos factores pueden influir en la capacidad de una persona para manejar situaciones estresantes de manera eficaz.

Por otra parte, los niveles de cortisol en sangre son un indicador biológico crucial para medir el estrés. El cortisol, según Zschimmer & Schwarz (2023), es una hormona glucocorticoide, una de las hormonas que se encarga de la regulación del metabolismo de carbohidratos favoreciendo la formación de glucosa y suprimiendo la actividad del sistema inmunológico (NLM). Un nivel elevado de cortisol en sangre puede ser

indicativo de un estado de estrés crónico, lo que puede tener implicaciones significativas para la salud física y mental (Blakemore, 2024).

En este trabajo, se emplearán métodos estadísticos avanzados para desarrollar modelos predictivos que integren tanto las características socio-económicas como los niveles de cortisol. Estos modelos buscarán identificar patrones y correlaciones que permitan predecir los niveles de estrés con precisión.

A través de un análisis exhaustivo de datos y la aplicación de metodologías rigurosas, se espera contribuir al entendimiento de los factores que influyen en el estrés en mujeres en el climaterio y ofrecer herramientas útiles para su prevención. Este periodo del climaterio se extiende en torno a 10 o 15 años y trata el proceso de transición antes y después de la menopausia (Torres et al., 2018) . Por último, este estudio pretende aportar conocimientos valiosos que puedan ser utilizados por profesionales para mejorar la calidad de vida de las mujeres en esta etapa de la vida.

1.2. Motivación del proyecto

Hoy en día, una gran parte de la población vive con situaciones de estrés de forma cotidiana, y esto afecta directamente a la salud física y mental (Blakemore, 2024). Mediante este estudio, se pretende encontrar qué variables pueden estar directamente relacionadas con los altos niveles de cortisol en sangre, como un instrumento para cuantificar los niveles de estrés, y por ello, la relación directa entre el cortisol y el estrés.

El tema escogido surgió gracias a la recomendación de mi tutor, quien me presentó varios temas en los que él estaba trabajando. Esta materia en ese momento me llamó mucho la atención, puesto que considero que cada vez más el estrés ha incrementado su relevancia en la sociedad, y además, me parece muy interesante indagar en los diferentes aspectos en los que que pueden estar influenciando.

Finalmente, al tratarse de los proyecto INMA y PAPILONG, que se están llevando a cabo actualmente, me motiva poder aportar información que sea relevante para el estudio y empezar un proyecto que facilite la comprensión y el acceso a los datos mediante alguna herramienta tecnológica.

1.3. Objetivos

El objetivo general de este TFG, es determinar los factores socio-económicos, los hábitos de salud y los comportamientos reproductivos y sexuales que se relacionan con los niveles de estrés de mujeres en etapa de climaterio comprendidas entre 40 y 70 años de la provincia de Valencia. Asimismo, se han definido una serie de objetivos específicos que se espera cumplir al concluir este proyecto. Estos se detallan a continuación:

I. Analizar y extraer los factores socio-económicos principales que influyen en los niveles del cortisol.
II. Construir y probar modelos predictivos utilizando técnicas estadísticas avanzadas.
III. Evaluar la efectividad de los modelos creados, analizando su capacidad de predicción de niveles de estrés en diferentes grupos de mujeres.

1.4. Relación con las asignaturas

En cuanto a la relación con las asignaturas, para llevar a cabo este análisis ha sido necesario recordar varias materias que he cursado a lo largo de la carrera universitaria.

En gran parte, este trabajo está relacionado con las asignaturas de estadística que he realizado en los diferentes cursos, como podrían ser Estadística, Métodos Estadísticos en Economía y por último Econometría. Mediante estas tres asignaturas, he podido construir una sólida base de conocimientos estadísticos hasta el día de hoy, fundamentales para la realización de este trabajo.

Para empezar, Estadística fue la primera asignatura que cursamos. En ella, se abordaron temas esenciales como la estadística descriptiva, introducción a la inferencia estadística y la presentación de las diferentes distribuciones de probabilidad. Por otra parte, en Métodos estadísticos en Economía pudimos explorar en detalle la inferencia estadística. Se estudió la inferencia para medias, proporciones, tablas de contingencias, y también las pruebas no paramétricas, entre otras. Para finalizar esta asignatura, se comenzó a dar importancia al análisis de la varianza y a los modelos de regresión, los

cuales junto con las pruebas no paramétricas serán claves para el desarrollo de este trabajo. En último lugar, tenemos la Econometría. Esta materia, sin duda es la más importante para poder realizar los modelos de regresión que tenemos planteados, así como para realizar modelos de series temporales. En esta última, se profundiza en conocer los modelos, aprendiendo cuándo es necesario aplicarlos y cómo interpretar sus resultados adecuadamente.

En cuanto a asignaturas no relacionadas con la estadística, debemos tener en cuenta también otras como Economía Española o Investigación Comercial. En la primera, se indaga en conocimientos fundamentales de la economía de nuestro país, pudiendo relacionarla con los factores socio-económicos que vamos a tratar, y también, nos proporciona un amplio conocimiento sobre los Objetivos de Desarrollo Sostenible, los cuales son necesarios para realizar un estudio que aporte valor a la humanidad. Por otra parte, mediante Investigación Comercial, profundizamos en la comercialización de un producto, donde tuvimos que llevar a cabo un estudio usando R, principal herramienta utilizada para analizar los datos que obtengamos.

1.5. Orden documental

En esta sección, se detalla la estructura y los temas que se abordarán en cada capítulo del presente trabajo. El objetivo es proporcionar una guía clara al lector para facilitar la comprensión del contenido.

En el primer capítulo, llamado “Introducción” se ofrece una breve presentación del trabajo, explicando las motivaciones que llevaron a su realización. Además, se concretan los objetivos generales y específicos que se busca alcanzar. Por último, en esta sección se ha llevado a cabo una revisión de las distintas asignaturas cursadas a lo largo de la carrera, y que guardan relación con el tema del TFG.

En el segundo capítulo, denominado “Marco teórico”, se exponen los conceptos fundamentales para el entendimiento de este estudio. En este, se explica a fondo qué es el cortisol, así como la forma en la que puede afectar a los niveles de estrés. Además, se exponen los dos proyectos de los cuales se van a obtener los datos con los que se va a trabajar.

Por otra parte, el capítulo principal es el de la Metodología. Aquí se detallan todos los pasos y análisis realizados para construir el modelo final. En primer lugar, se explica cómo se han obtenido los datos, se limpian y procesan y por último, se detalla la organización de la base de datos así como las variables que la forman. Por otra parte, se lleva a cabo un análisis individual para comprender las variables por separado y su relación con los niveles de cortisol en sangre, así como un análisis en profundidad sobre las variables dependientes. Por último, se explican los procedimientos e información necesaria para comprender cómo se realiza la construcción del modelo.

En el capítulo de “Análisis y Discusión de los resultados” se presentan las métricas obtenidas, relacionándolas con el marco del trabajo y evaluando los resultados. En esta sección, se desarrollan varios modelos de regresión múltiple hasta encontrar el modelo más adecuado, basándonos en los criterios previamente establecidos y el cual cumpla con los requisitos de validez. Finalmente, se realiza una interpretación del modelo y la predicción del cortisol en sangre que puede tener una persona dependiendo de los valores que tengan las variables que hayan salido significativas.

Por último, nos encontramos con las Conclusiones, donde se resume todo el trabajo realizado, así como los resultados a los que se ha llegado y los objetivos que se han cumplido. También, se muestran tanto las limitaciones y dificultades encontradas, así como los logros obtenidos y recomendaciones sobre posibles líneas de trabajo futuro.

Al final del trabajo se encontrará la bibliografía con todas las fuentes aplicadas y los anexos donde se encuentra el script de código R empleado y la relación de este TFG con los Objetivos de Desarrollo Sostenible.

2. Marco teórico

Para entender este trabajo, primero hay que poner en contexto y explicar ciertos términos y su funcionamiento. En primer lugar, es imprescindible saber qué es el cortisol y cómo sabemos si los niveles que tiene una persona son elevados o si por el contrario, se encuentran dentro de la normalidad. El cortisol es una hormona de tipo glucocorticoide que afecta a los sistemas corporales y que se encarga de la regulación del metabolismo de carbohidratos (Levine et al., 2006). También es conocida como la hormona del estrés, ya que su producción aumenta en respuesta al estrés. Para saber los niveles de cortisol de una persona, los médicos pueden solicitar diferentes tipos de pruebas de cortisol como pueden ser los niveles de cortisol en sangre, en el pelo, en la orina o en la saliva (NLM). En este trabajo, nos centraremos en los niveles de cortisol en sangre y en pelo.

Por otra parte, unos altos niveles de cortisol pueden llevarnos a padecer enfermedades como el Síndrome de Cushing debido a una exposición crónica a un exceso de cortisol (Lethielleux, 2020). Es por ello que hay que tener un control sobre nuestros niveles de cortisol. También es cierto que los niveles de cortisol pueden variar dependiendo del momento del día en el que se realicen las pruebas ya que la secreción de cortisol durante el día no es homogénea (Quiceno et al., 2016). En conclusión, los niveles de cortisol se consideran:

Tabla 1: *Niveles de cortisol*

Fuente: *Elaboración propia con información extraída de Quiceno et al. (2016) y Dressl et al. (2018) para el cortisol en sangre; Faresjö et al. (2024) para el cortisol en pelo*

	En sangre	**En pelo**
Cortisol bajo	< 5 µg / dL ≈ 138 nmol/L	< 17 mg/pg ≈ 4,7 nmol/L
Cortisol normal	7 µg / dL - 25 µg / dL ≈ 193,13 nmol/L - 690 nmol/L	20 mg/pg - 80 mg/pg ≈ 5,5 nmol/L - 22 nmol/L
Cortisol alto	> 25 µg / dL ≈ 690 nmol/L	> 84 pg/mg ≈ 23,18 nmol/L

También es importante hablar sobre el climaterio. Climaterio es un término derivado del griego "klimakter" y se refiere a un periodo natural de transición y adaptación a cambios biológicos, psicológicos y sociales. Este proceso de transición se extiende durante años antes y después de la menopausia, debido al agotamiento ovárico. Abarca el paso de la edad reproductiva a la no reproductiva, marcado por una disminución en la producción de estrógenos y la eventual pérdida de capacidad de los ovarios para producir hormonas, folículos y ovocitos. Así, el climaterio incluye la perimenopausia, que es la etapa de transición hacia la menopausia (indicativa del fin del periodo reproductivo), la menopausia misma, que es el cese permanente de la menstruación, y la postmenopausia. Este proceso suele durar aproximadamente entre 10 y 15 años (Torres et al., 2018; Taechakraichana, 2002).

En cuanto a los proyectos que se están llevando a cabo, nos encontramos con el proyecto INMA - Infancia y Medio Ambiente y el proyecto PAPILONG. Estos proyectos tienen un diseño transversal y cuentan con la participación de 381 mujeres entre los dos, con un rango de edad entre los 40 y los 65 años. Las participantes pertenecen a dos cohortes distintas cuya incorporación se produjo entre junio del año 2019 y diciembre del 2022.

El primero es un estudio analítico observacional longitudinal de base poblacional. La muestra seleccionada es un subconjunto de mujeres pertenecientes a la cohorte INMA-Valencia que originalmente empezó con 855 participantes entre 2003 y 2004 durante su embarazo. Tanto las mujeres como sus correspondientes hijos han sido seguidos a lo largo de los años, realizando una nueva visita entre 2019 y 2021. De las 855 madres iniciales se cuenta con datos de 249 madres que participaron en la visita de 2019. En esta visita, se recogieron muestras de su microbiota vaginal acompañada de información sociodemográfica, estilo de vida, comportamiento sexual y biológica rellenada por ellas mismas, además de la medida de síndrome de Burnout mediante escala MBI. Cabe destacar que la mayoría de las participantes contaban con una clase social media o media-alta.

En cuanto al segundo proyecto (https://links.uv.es/0ZFlslW) está compuesto por una cohorte de 132 mujeres. En este caso, la mayoría se encontraban en riesgo de pobreza y/o exclusión social ya que muchas de ellas eran trabajadoras sexuales, personas sin hogar o mujeres que buscan ayuda en bancos de alimentos, todas ellas

participantes de distintas ONG de Valencia, como podría ser “Por ti Mujer”, “Cruz Roja Española” o “Amigos de la Calle”, entre otras. Esta segunda cohorte fue formada entre los años 2021 y 2022.

Para este TFG se utilizarán los datos del cortisol en pelo y en sangre obtenidos en la primera visita, así como otras variables socio-demográficas y biológicas que se explicarán en la siguiente sección de ambos proyectos, INMA y PAPILONG.

Finalmente, el proyecto en el que se enmarca este estudio ha obtenido la aprobación del Comité de Ética de la Conselleria de Sanitat Valenciana (DICTAMEN CEI-SP 20220708/061). Además, todas las participantes firmaron un consentimiento informado antes de completar las encuestas y la toma de muestras, y recibieron una explicación detallada sobre los objetivos del estudio, los riesgos y procedimientos involucrados, la cobertura del seguro, su derecho a retirarse en cualquier momento y la forma en la que se comunicarán los resultados finales del estudio.

3. Metodología

3.1. Obtención de la base de datos

En cuanto a la base de datos, se han utilizado los datos de los proyectos anteriormente mencionados, el proyecto INMA y el proyecto PAPILONG. Como ya se ha comentado, está compuesta por 381 mujeres divididas en dos cohortes cada una perteneciente a uno de los proyectos. Para facilitar el análisis y unificar la información de ambas cohortes se realizó una combinación de las dos bases de datos originales.

Durante este proceso, se añadió una columna adicional que indica la base de datos inicial a la que pertenece cada participante, asignando un “0” a las personas del proyecto INMA y un “1” para el proyecto PAPILONG. De esta forma, se puede distinguir fácilmente entre las participantes de cada cohorte en la nueva base de datos.

Además, las bases de datos no estaban completas, sino que era necesario relacionarlas con bases de datos adicionales que contenían los niveles de cortisol en sangre y en pelo de cada participante, siendo necesario combinarlas mediante los identificadores que se les había asignado a cada una de ellas.

Finalmente, al unir las dos bases de datos en R, las columnas comunes en ambas, como por ejemplo AROPE_cat, se duplicaron con nombres distintos, es decir, para el proyecto INMA se creó como AROPE_cat.x y para el proyecto PAPILONG se creó como AROPE_cat.y. Por lo tanto se tuvieron que unificar en una nueva columna que fuera AROPE_cat conservando el valor de la variable en una única columna. Esta reorganización permite que se pueda trabajar de una forma más fácil y eficiente manteniendo toda la información relevante de manera combinada.

3.2. Presentación de las variables dependientes

Para entender el trabajo que se va a realizar tiene que quedar muy clara la diferencia entre las variables independientes o variables predictoras y las variables dependientes o variables respuesta. En el caso de las variables dependientes su valor se verá afectado en función de los valores que adopten las variables independientes. Es decir, se intentará predecir o explicar los cambios que se van a producir en la variable respuesta, en función de la influencia que tengan las variables predictoras. Por otra parte, las variables independientes o predictoras describen los factores que pueden afectar a la variable dependiente y cómo afectaría a esta última su valor en función del cambio que se produzca en las primeras.

Por ello, empezamos explicando las variables respuesta. Se quiere estudiar las variables del cortisol en sangre y el cortisol en pelo. La primera variable refleja el cortisol en tiempo real y en un momento puntual. Sin embargo, el cortisol en pelo proporciona un promedio de los niveles de cortisol acumulado a largo plazo, entre varias semanas y meses, convirtiéndolo en un marcador de estrés crónico (NLM). Mediante los siguientes histogramas, podemos visualizar la existencia y la diferencia entre los niveles de cortisol en sangre a corto plazo, y los niveles de cortisol en pelo a largo plazo de nuestra muestra de INMA y PAPILONG:

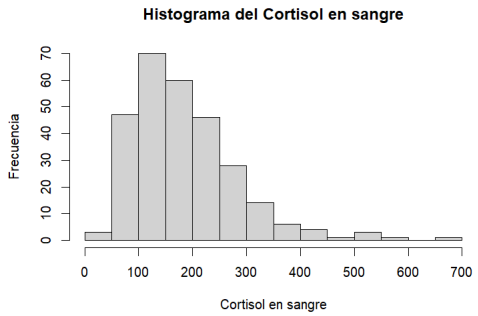


Gráfico 1: *Histograma del cortisol en sangre*
Fuente: *Elaboración propia*

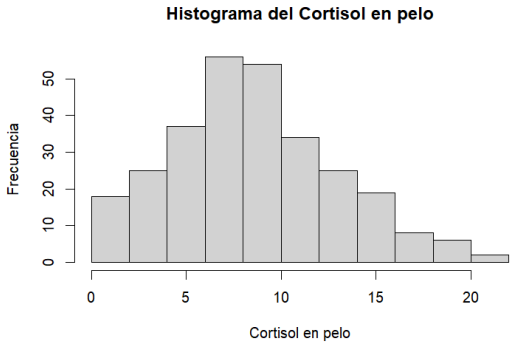


Gráfico 2: *Histograma del cortisol en pelo*
Fuente: *Elaboración propia*

Tal y como se observa en los gráficos, la mayoría de las participantes se encuentran con un nivel de cortisol en sangre entre 50 y 200 nmol/L, y en el caso de cortisol en pelo se encuentran entre 4 y 12 nmol/L. Como ya se ha mencionado

anteriormente, estos niveles no serían preocupantes aunque tampoco se podrían considerar como un nivel bajo de cortisol.

Por otra parte, si comparamos los niveles de cortisol en sangre, con los niveles de cortisol en pelo se observa que existe una relación negativa entre ambos. El valor de la correlación entre estas dos variables es de -0,1006456, lo que indica que no hay una relación lineal fuerte ya que el valor es cercano a 0 y también, sugiere una relación inversa. Este valor puede explicar que a medida que los niveles de cortisol en sangre aumentan, los del cortisol en pelo disminuyen ligeramente. Asimismo, la gran cantidad de puntos dispersos puede también indicar que la correlación entre las dos variables es débil.

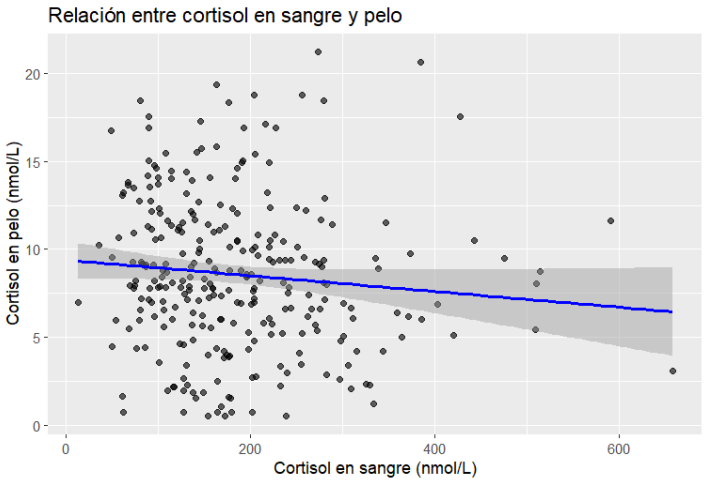


***Gráfico 3**: Gráfico de dispersión entre el cortisol en sangre y en pelo*
***Fuente**: Elaboración propia*

En conclusión, mediante estos gráficos hemos podido deducir que los niveles de cortisol en sangre cubren un intervalo más amplio llegando a tener los valores más elevados. Y los valores del cortisol en pelo están más concentrados en un rango intermedio. Además, la relación entre las dos variables no es muy fuerte, indicando que los valores de cortisol en sangre son claramente superiores a los del cortisol en pelo.

3.3. Presentación de las variables independientes

En la siguiente tabla, se presentan las diferentes variables independientes o predictoras que vamos a estudiar. En ella se incluyen el nombre que le hemos dado en la base de datos, una pequeña descripción de lo que indica y finalmente el tipo de variable que es:

Tabla 2: *Resumen de las variables de la base de datos*

Fuente: Elaboración propia

VARIABLE	NOMBRE EN LA BASE DE DATOS	DESCRIPCIÓN	TIPO
Age	Age	Indica la edad de la participante	Cuantitativa
Education level	Education	La educación recibida de la participante	Cualitativa (primary studies, secondary studies y university studies)
Work	Work	Si está trabajando actualmente	Cualitativa (No, Si)
Social Class	SC	Indica la clase social de la participante	Cualitativa (CS I+II, CS III, CS IV+V)
Perceived Health Status	Health_Stat	Estado de salud percibido durante el último año	Cualitativa (Very good, Good, Fair-bad-very-bad)
Smoking	Smoking	Saber si fuma o no	Cualitativa (No, Si)
Alcohol intake (Alcohol)	Alcohol	Saber si bebe alcohol	Cualitativa (No, Si)
Body Mass Index (BMI)	BMI	Mide el contenido de grasa corporal en relación a la estatura y el peso (IMC)	Cualitativa (< 25 kg/m2, 25-29,9 kg/m2, >= 30 kg/m2)
Menopause status	Menop_stat	Descripción del climaterio de la mujer	Cualitativa (No menopause, Menopause)
Age 1ª sexual intercourse (years)	Intercourse	Edad con la que mantuvieron la primera relación sexual	Cuantitativa
Current partner	Cpartner	Si tiene pareja actualmente	Cualitativa (No, Si)
Last 12 months n of partners	partner12	Nº de parejas en los últimos 12 meses	Cualitativa (0, 1, >1)
Oral contraceptives (Oral)	Oral	Si toma o ha tomado anticonceptivos en el último año	Cualitativa (No, Si)

VARIABLE	NOMBRE EN LA BASE DE DATOS	DESCRIPCIÓN	TIPO
Intrauterine device	IUD	Si usa o ha usado un DIU en el último año	Cualitativa (No, Si)
Preservatives	Preservat	Si utiliza preservativo	Cualitativa (No, Si)
Antibiotics	Antib	Si consume algún antibiótico	Cualitativa (No, Si)
Antifungal and anti urinary tract infections	Antif_Antiu	Si consume o ha consumido algún tratamiento para infecciones en el último año	Cualitativa (No, Si)
Probiotics	Prob	Si ha consumido probióticos en el último año	Cualitativa (No, Si)
Workout	Workout_2c	Tiempo dedicado a hacer deporte en su tiempo libre	Cualitativa (Less than 3 hours/week, 3 or more hours/week)
Workout perceived	Workout_perciv	Estilo de vida en cuanto al deporte	Cualitativa (Sedentary/Low active, Moderate active , Active/Very active)
Negative hygiene products	MicroAlterNeg	Si utilizan productos de higiene íntima que afecten de manera negativa a su salud	Cualitativa (No, Si)
Positive hygiene products	MicroAlterPos	Si utilizan productos de higiene íntima que afecten de manera positiva a su salud	Cualitativa (No, Si)
Country_2catSpain	Spain	De donde son	Cualitativa (No, Si)
Menstrual cycle	M_Cycle	En que fase del ciclo menstrual se encuentran	Cualitativa (None, The follicular phase, Ovulation, The luteal phase)
Vaginal micriobiota collection data	V_Date	Fecha en la que se recolectaron los datos de la microbiota en sangre	Cuantitativa
Date of last menstrual period	M_Date	Fecha del último periodo	Cuantitativa
At risk of poverty or social exclusion (AROPE)	AROPE	Si se encuentra en situación AROPE	Cualitativa (No, Si)
Base de datos	BBDD	Para saber la base de datos inicial	Cualitativa (0, 1)

3.4. Preprocesamiento

3.4.1. Datos anómalos

Para realizar un buen preprocesamiento de datos hay que analizar los datos anómalos o valores atípicos. Estos son observaciones dentro de un conjunto de datos que se desvían significativamente de un patrón general y pueden influir negativamente en los resultados al aplicar métodos estadísticos (Luis Marcano , Wilmer Fermín, 2013)

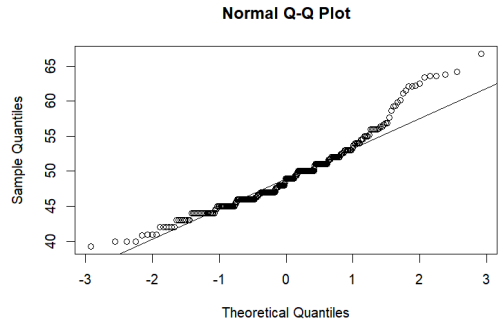


Gráfico 4: *Qqplot variable Edad*

Fuente: *Elaboración propia mediante R*

En cuanto a la “Edad de la primera relación sexual”, podemos observar que la mayoría de valores se encuentran entre los 17 y los 21 años, siendo la media de 19 años. En este caso, los valores más bajos se podrían aceptar si tenemos en cuenta que una parte de la población mantiene su primer encuentro sexual en edades muy tempranas, así como los valores por encima de 25 años que también pueden ser considerados.

En estos casos, tan solo se pueden analizar las variables cuantitativas, ya que las cualitativas disponen únicamente de las opciones que se han predeterminado antes de realizar la encuesta, por ello podemos observar las variables “Edad” y “Edad de la primera relación sexual". Para la primera, podemos deducir mediante el siguiente gráfico que los únicos valores anómalos podrían ser 39 y 70 años, pero como se pueden mantener en el estudio, se decide no eliminarlo.

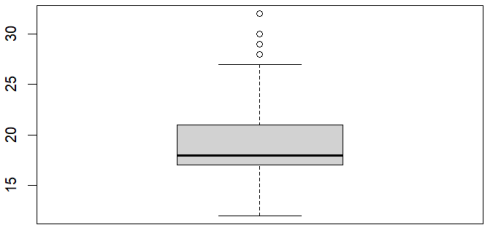


Gráfico 5: *Boxplot variable Edad 1ª relación sexual*

Fuente: *Elaboración propia con R*

3.4.2. Valores faltantes

Para saber cuales son los valores faltantes, primero se ha realizado un recuento del número de “not available” (NA) que hay en cada columna para considerar eliminar las que tengan un porcentaje alto de datos sin contestar. Si nos fijamos en la base de datos para el proyecto PAPILONG faltan las variables de la “Clase Social”, “Deporte percibido" y las “Observaciones". Después de realizar un análisis, se decide eliminarlas ya que en cuanto a la “Clase Social” aunque pueda ser significativa para este estudio, tenemos la variable “AROPE” que puede actuar como un indicador de la clase social; para la variable del “Deporte percibido” queda cubierto con la “Cantidad de deporte realizado por los participantes” y las “Observaciones” son mayormente sobre el estado de la menopausia que queda demostrado en la variable del “Estado de la menopausia”. Aparte de estas tres, la única variable que también se podría eliminar sería la variable “Drinking” ya que tiene un 9,09% de valores sin contestar. Sin embargo al no alcanzar el 20% de datos sin rellenar se podrían valorar otras opciones como la imputación de valores rellenando los huecos faltantes con estimaciones, o la eliminación de las filas que no contengan valor.

Para este tipo de estudio se pueden elegir varias opciones a la hora de decidir qué se puede hacer con los valores faltantes. Por una parte, se puede realizar la imputación de los datos, en la cual mediante el paquete “mice” en R conocido para realizar la imputación multivariante, se añaden valores aleatorios a la base de datos en todos los valores nulos o NA, teniendo en cuenta los valores entre los que se puede escoger para cada columna (Medina y Galván, 2007). Por otra parte, tenemos la eliminación de los datos por lista completa. Esta opción implica eliminar todas las filas que tengan uno o más valores faltantes. Gracias a esta posibilidad, los datos que se tratan son más fiables ya que no se han extraído mediante valores aleatorios pero hay que tener en cuenta que se puede reducir excesivamente el tamaño de la muestra. A continuación, se adjunta una tabla con los valores faltantes de cada variable en número y porcentaje:

Tabla 3: *Cantidad de valores faltantes por variable*

Fuente*: Elaboración propia*

Idnum	0	**Age**	0	**Education**	3	**Work**	2	**Health_Stat**	1	**Smoking**	2
Drinking	35	**BMI**	0	**Menop_stat**	0	**Intercourse**	0	**Cpartner**	2	**partner12**	22
Oral	1	**IUD**	1	**Preservat**	1	**Antib**	4	**Antif_Antiu**	4	**Prob**	4
Workout_2c	6	**MicroAlterNeg**	2	**MicroAlterPos**	4	**Spain**	0	**M_Cycle**	0	**AROPE_cont**	32
AROPE_cat	5	**Cortisol_sang**	0	**Cortisol_pelo**	0	**V_date**	0	**M_Date**	0	**BBDD**	0

Tras varias pruebas, se decide finalmente optar por la opción de eliminar determinadas filas de datos, ya que aunque se supriman unos 90 valores aproximadamente, la base de datos sigue teniendo suficientes valores para ser significante, y además, nos aseguramos de que los valores con los que vamos a trabajar son fiables.

Con respecto a los valores duplicados de la variable “Cortisol_pelo”, se debe cambiar todos aquellos valores en los que aparezca “<1,5” por el valor “0,75” ya que cuando son valores muy pequeños la precisión se pierde. De hecho, se ha comprobado que estos valores son siempre menores a “1,5” por lo que decidimos optar por un valor promedio de “0,75” que se encuentra entre “0” y “1,5”.

Además, hay que comprobar que los valores coincidan en las dos bases de datos y realizar los cambios necesarios para que puedan congeniar. Al empezar a trabajar con las dos bases de datos, se pudo observar que no se estaban aplicando las respuestas con los mismos valores, es decir, para una de las bases las variables cualitativas con varias opciones empezaban con el número 0 al pasarlas a cuantitativas. Sin embargo, para la otra base de datos comenzaba con el 1. Por ello, se tuvo que corregir los valores de la segunda para que estuvieran en línea y coincidieran los datos.

Por último, al basar nuestro estudio en la cantidad de cortisol en sangre, las participantes que finalmente no hayan obtenido resultados válidos en las pruebas, ya

sea por la ausencia de realización o resultados defectuosos, también serán excluidas de la muestra.

3.5. Análisis de las variables

Tras la tabla resumen de las variables y el preprocesamiento de los datos, se realizará un análisis general de las variables exponiendo sus características y relación con la variable independiente:

1. **Edad**: variable cuantitativa numérica que nos informa sobre las edades de las participantes para este estudio. Mediante esta variable observamos que tiene un mínimo de 39 años y máximo de 70 años, con una media de 52 años y encontrándose la mayoría de participantes entre los 45 y 55 años.
2. **Educación**: variable cualitativa que nos demuestra el nivel de estudios que tienen las participantes del trabajo. Está dividida en diferentes opciones que son: *primary studies*, *secondary studies* y *university studies*. La mayoría se encuentran en la franja de estudios secundarios con un 43%.
3. **Trabajo:** variable cualitativa que nos indica si la participante está trabajando o no en el momento en el que se hizo la encuesta. En este caso, se observa una gran mayoría de mujeres que trabajan, llegando el 76% siendo únicamente un 24 % de desempleadas
4. **Estado de Salud Percibido:** se encuentra dividido en “Muy bueno”, “Bueno” y “Regular-malo” y “Muy malo” siendo el 55,63% “Bueno”.
5. **Tabaco:** la siguiente variable muestra si la participante fuma o no, quedando un total de 213 mujeres que no fuman, frente a 71 que sí que lo hacen.
6. **Alcohol:** esta variable muestra si toma bebidas con alcohol, en este caso el número aumenta considerablemente, quedando con un 53,87% las que sí que beben, es decir, más de la mitad.
7. **Índice de Masa Corporal**: esta variable cualitativa se pregunta en la encuesta para saber si se encuentran con valores menores a 25 kg/m2, si están entre los 25 - 29 kg/m2 o si se encuentran por encima de 30 kg/m2. En este caso, la mayoría sería para la primera opción, es decir, un 39,8% con un IMC < 25 kg/m2.
8. **El estado de la menopausia:** esta variable cualitativa es crucial en este estudio mostrando que tan solo un 28% se encuentran en este estado, para poder abarcar todo el climaterio.

9. **Edad de la primera relación sexual:** esta variable cuantitativa nos informa sobre las edades en la que las participantes mantuvieron su primera relación sexual, teniendo un mínimo de 12 años y un máximo de 32, con una media de 19 años.
10. **Pareja**: también es importante saber si se encontraban con una relación de pareja en el momento en el que se hizo la muestra mostrando el 77,5% que sí, osea, una gran mayoría.
11. **Nº parejas en los últimos 12 meses:** por otra parte, de todas las participantes, el 80,3% tuvo tan solo 1 pareja durante los últimos 12 meses.
12. **Métodos anticonceptivos**: el 56% de las participantes toma pastillas anticonceptivas de forma oral, y además el 23,6% ha usado el DIU durante el último año.
13. **Uso de preservativo**: también es importante comentar el uso de preservativo que aunque sea mayoritario, tan solo el 67% lo usa de forma habitual.
14. **Antibióticos**: en cuanto a la ingesta de antibióticos tan solo un 25% sí que los toman de forma diaria.
15. **Antifúngicos**: esta variable cualitativa nos muestra si han tomado algún tratamiento para infecciones durante el último año, mostrando que tan solo 51 mujeres, es decir el 18%, sí que lo han hecho.
16. **Probióticos**: esta es una variable cualitativa que nos indica si ha consumido probióticos en el último año y tan solo un 7,4% sí que los tomaron.
17. **Deporte:** esta variable cualitativa demuestra la cantidad de mujeres que realizan deporte de forma regular y en más o menos de 3 horas por semana, mostrando que la mayoría con un 57% no llega a realizar 3 horas de deporte por semana.
18. **Productos íntimos que afectan negativamente**: para esta variable, se demuestra si los productos que han usado han afectado negativamente a su salud íntima, respondiendo con “Sí” un total de 119 mujeres, es decir, el 41%.
19. **Productos íntimos que afectan positivamente:** por el contrario a la anterior, esta variable demuestra si han usado productos que han afectado de manera positiva a su salud íntima. Tan solo 20 mujeres, es decir un 7% del total, han afirmado que sí que han encontrado productos positivos para su salud íntima.
20. **Origen**: en cuanto al origen, nos muestra si son extranjeras o si son españolas. Un 72,5% de las participantes afirmó que son de origen español.
21. **Ciclo menstrual:** esta variable cualitativa nos muestra en qué momento del ciclo menstrual se encuentran, cosa que en un principio se puede pensar que afecta

en gran parte. Está dividida en 4 fases: ninguna, fase folicular, ovulación y fase lútea. La fase folicular es aquella que ocurre entre el primer día del sangrado hasta 14 días después y en ella el endometrio empieza su fase proliferativa con aumento del espesor de los vasos, estroma y estructuras glandulares. La ovulación se produce tras la fase folicular y en ella comienza la formación de cuerpo lúteo. Finalmente, la última fase es la fase lútea que comprende el tiempo entre la ovulación y la menstruación, en esta fase el endometrio empieza su étapa secretora (Jiménez y Aguilá, 2017). Después de esta explicación, un 49% de las mujeres no se encontraban en ninguna fase, ya sea por menopausia o por menstruación, un 15% en fase folicular, un 10% en ovulación y finalmente un 26% en fase lútea.

22. **AROPE (At risk of poverty or social exclusion)**: la siguiente variable cualitativa es un indicador que combina tres conceptos: baja intensidad laboral de hogar, privación material severa y pobreza. La situación de baja intensidad laboral se cumple cuando el cociente entre los meses trabajados por todos los miembros de hogar y los meses que podrían haber trabajado es inferior a 0,2 considerando los adultos de 18 a 59 años exceptuando los estudiantes de 18 a 24 años. La privación material severa se refiere a hogares que no pueden permitirse 4/9 ítems básicos, incluyendo retrasos en pagos relacionados con la vivienda, incapacidad para afrontar gastos imprevistos, no poder permitirse vacaciones, no consumir carne o pescado cada dos días, no mantener una temperatura adecuada en la vivienda y carecer de bienes como teléfono, automóvil, televisor en color o lavadora (Lechuga, Luque y Martínez, 2015). Finalmente, la pobreza se define por ingresos inferiores al 60% de la renta media nacional, estableciendo el umbral de pobreza. La población que cumpla al menos uno de estos tres criterios será incluida en la población AROPE. Para este estudio, un total de 109 mujeres, un 38,4% se encuentran dentro de la población AROPE, por cumplir al menos uno de los criterios mencionados.

Tras esta explicación y para poder observar de manera gráfica lo que se ha comentado, se muestra una tabla con boxplots de las variables objetivas de estudio (cortisol en sangre y en pelo) por cada una de las variables cualitativas anteriormente mencionadas:

Tabla 4: *Gráficos de boxplots de todas las variables cualitativas frente al cortisol en sangre*
Fuente: *Elaboración propia*

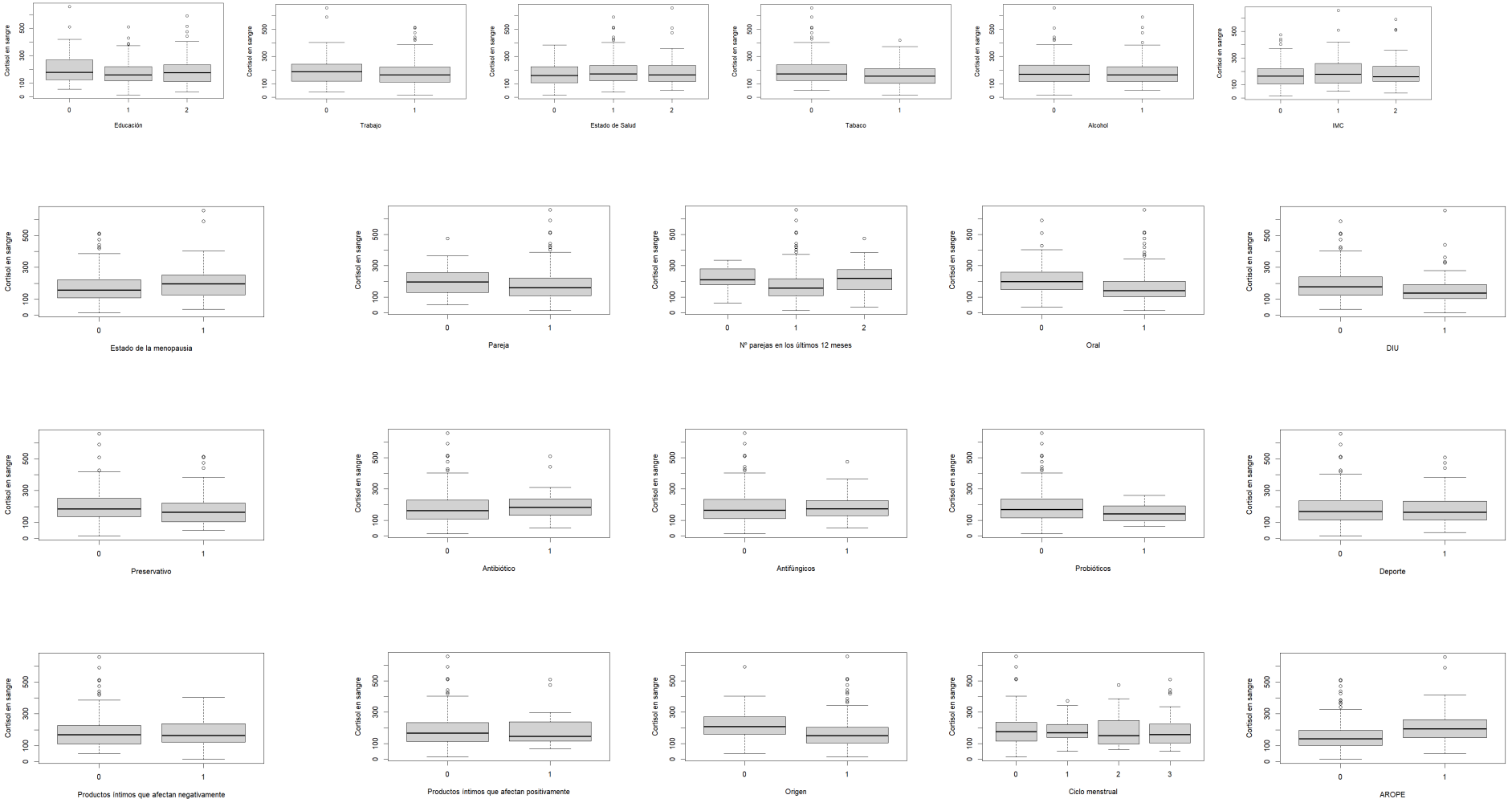
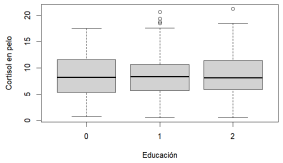
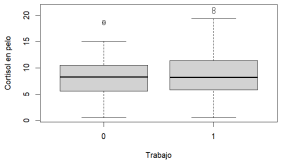
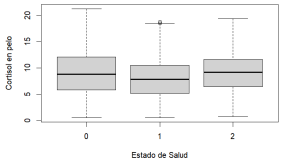
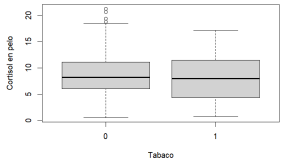


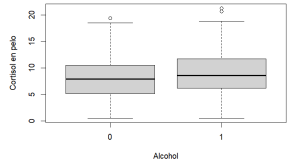
Tabla 5: *Gráficos de boxplots de todas las variables cualitativas frente al cortisol en pelo*
Fuente: *Elaboración propia*

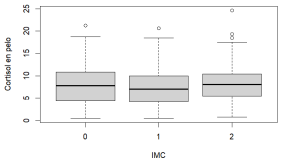


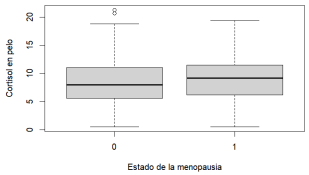


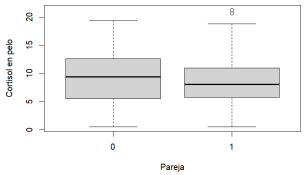


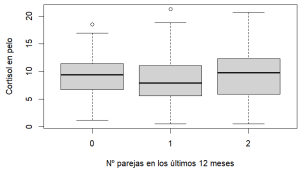


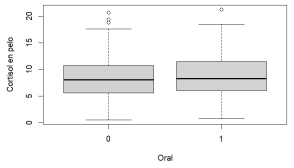


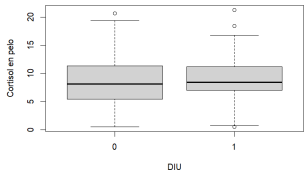


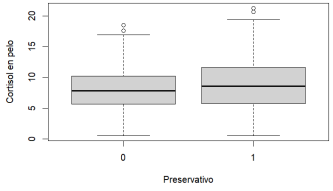


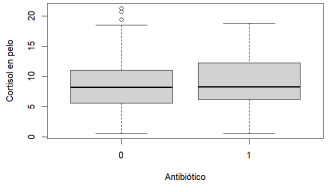


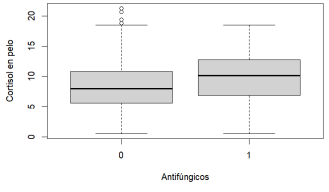


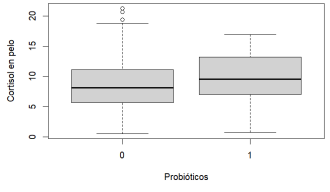


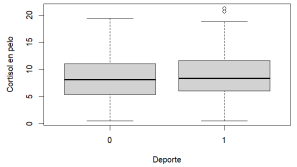


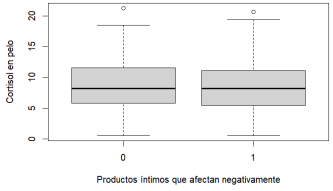


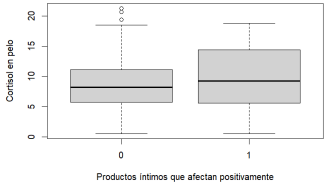


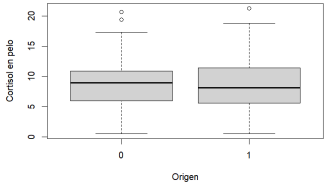


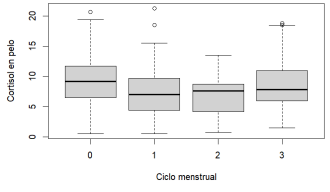


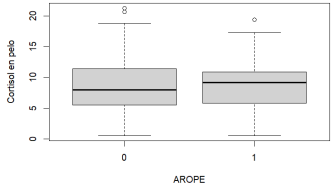












3.6. Procedimientos e información para la construcción del modelo

Una vez se ha terminado la limpieza de la base de datos, se procede a trabajar con esta para sacar modelos de regresión lineal y validarlos, hasta llegar a los modelos que resulten adecuados.

3.6.1 Regresión lineal

Para comenzar, teniendo en cuenta que nuestro objetivo es cuantificar cómo cada una de estas variables influye en los niveles de cortisol, lo haremos mediante un modelo de regresión lineal, ya que permite medir el efecto que tiene cada variable independiente sobre la variable dependiente manteniendo las demás variables constantes.

La regresión lineal es una herramienta estadística utilizada para examinar como una variable dependiente se relaciona con una o varias variables independientes. El objetivo principal es conseguir una línea de regresión que describa de la mejor forma posible cómo se modifica la variable dependiente, en función de los cambios en variables independientes. Es decir, muestra la relación que hay entre las variables y cómo se asocian entre ellas (Laguna, 2014; Granados, 2016). La ecuación básica de la línea de regresión de este tipo de modelo es:

$$Y = \beta 0 + \beta 1 X1 + \ldots + \beta p Xp + U$$

donde:

- Y representa la variable dependiente o respuesta, que es el valor que se intenta predecir.
- Xp es la variable independiente o explicativa, la cual se relaciona con Y.
- $\beta 0$ es el intercepto, es decir, el valor promedio de Y cuando Xp es cero.
- βp indica cuánto cambia el valor promedio de Y por cada unidad de incremento en Xp.
- U es el término de error o residuo, que recoge la variación en Y que no es explicada por Xp.

Además, este ajuste se realiza mediante el método de los mínimos cuadrados, el cual minimiza la suma de los errores al cuadrado entre los valores observados y los predichos, buscando que la línea sea lo más cercana posible a los valores (Laguna, 2014; Granados, 2016).

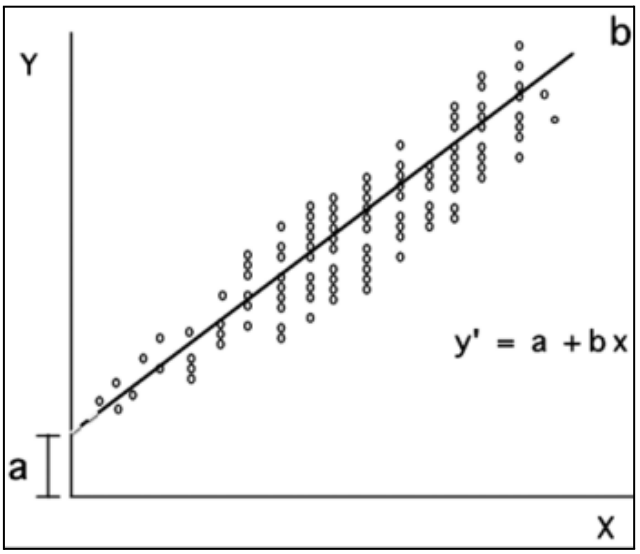


***Gráfico 6**: Ejemplo de regresión lineal*

***Fuente**: Dagnino, J. (2014). Regresión lineal.*

Mediante el gráfico anterior, como indica en el artículo de Dagnino, *"Regresión lineal*": la línea calculada con la ecuación es la que mejor representa los datos, donde "a" es la intersección de la línea con el eje de las ordenadas y "b" es la pendiente de la línea.

3.6.2 Selección de modelos

Para obtener las variables independientes significativas se pueden usar diferentes técnicas que construyan el modelo, como por ejemplo el método de regresión Stepwise con el paquete "MASS" ya que según Venables y Ripley (2013), este paquete con la función stepAIC, fue desarrollado para seleccionar automáticamente un subconjunto óptimo de variables independientes para nuestro modelo predictivo, el cual minimice el AIC mediante un sobreajuste. En este caso, disponemos de una base de datos con muchas variables, por lo que antes de construir el modelo se debe realizar una selección previa, por lo que este método es perfecto para ello. Para eliminar o introducir una variable en el modelo, se han utilizado dos criterios que se explican a continuación

En primer lugar, la regresión por pasos (stepwise regression) es un método utilizado para construir modelos de regresión de manera iterativa, es decir, añadiendo y eliminando variables con el objetivo de optimizar el ajuste del modelo. Lo que se busca es llegar a un modelo en el que solo se queden las variables que son estadísticamente significativas y eliminar aquellas que no aportan valor explicativo (Agostinelli, 2002) .

Además, existen tres enfoques principales dentro del stepwise regression: forward stepwise regression (hacia delante), backward stepwise regression (hacia atrás) y both stepwise regression (bidireccional hacia delante y hacia atrás). En el primer método nombrado, el modelo comienza sin variables y va añadiendo una por una, evaluándolas y solo se mantiene si es significativa. Esto se repetirá hasta que ya no quede ninguna variable más por probar. Por el contrario, la backward regression empieza incluyendo todas las variables y va eliminando aquellas que no son significativas hasta que se obtiene el mejor modelo. Finalmente, el método both regression combina ambos enfoques añadiendo y eliminando de manera simultánea las variables. Este modelo puede ser muy útil para optimizar el modelo sin la necesidad de probarlo manualmente paso por paso (Adam, H., Khadija, K., & Suzanne, K., 2022). Por todo ello, se utilizará el método **both regression** para este trabajo ya que permite simplificar el modelo añadiendo y eliminado las variables de forma simultánea, además de mejorar su precisión así como la capacidad de la generación de predicciones.

En segundo lugar, tenemos el criterio de información de Akaike (AIC) el cual se usa para evaluar y comparar modelos de regresión, sobre todo en procesos de selección de variables (Cavanaugh y Neath, 2019). En los métodos de selección como el stepwise en nuestro caso, el AIC ayuda a añadir o eliminar variables ya que se incluyen aquellas que reduzcan el AIC y se eliminan las que no mejoran significativamente el criterio optimizando de esta forma el modelo final.

Por otro lado, la significatividad de la variable también puede ser utilizada para la selección. El nivel de significatividad que usamos es el de 0,05 (Kennedy-Shaffer, 2019). Para lograrlo, en las funciones codificadas se han utilizado los valores de p_enter=0,06 y p_remove=0,1. Una de las razonas mediante las cuales se puede justificar es que da una mayor flexibilidad en la inclusión de variables que podrían ser significativas, puesto que al ampliar el umbral de inclusión permite que las variables que se encuentran entre el 0,05 y 0,06 puedan entrar en el modelo y aportar información

importante en la predicción. Además, también permite que el modelo se adapte de una mejor forma al contexto del estudio, esto quiere decir que con el umbral de 0,05 podría llegar a ser demasiado restrictivo, por lo tanto con el aumento de los dos valores permite adaptarse a la variabilidad de los datos. Por estas razones, se ajustan los valores de p_enter y p_remove para asegurarse que el modelo es lo suficientemente flexible como para incluir variables que son significativas pero evitar incluir las que son innecesarias.

3.6.3 Validación del modelo

Para que los modelos sean estadísticamente fiables y útiles para hacer predicciones, tanto para el modelo del cortisol en sangre como para el modelo del cortisol en pelo, se procede a realizar la validación de estos modelos para comprobar si cumplen con las hipótesis del modelo de regresión. Esto se hace demostrando que el error es ruido blanco (Venables y Ripley, 2013). Este modelo se usa como un modelo para describir errores o variaciones aleatorias. Si los residuos de un modelo se comportan como ruido blanco, se puede decir que son puramente aleatorios, por lo tanto no afectarán al modelo final. Los requisitos que debe cumplir un modelo para considerarse ruido blanco son:

- **Media cero:** que los residuos U tengan una media esperada de cero implica que no hay una tendencia hacia valores positivos ni hacia valores negativos.
- **Varianza constante:** también conocida como homocedasticidad, se observa si la varianza de los residuos es constante.
- **Incorrelación:** esto nos demuestra que los residuos de cada observación son independientes entre sí, es decir, que no existe correlación entre ellos.
- **Normalidad:** la normalidad nos permite saber si los residuos están lo suficientemente bien alineados para admitir normalidad y nos permite identificar valores anómalos.

4. Análisis y resultados

4.1 Construcción del modelo

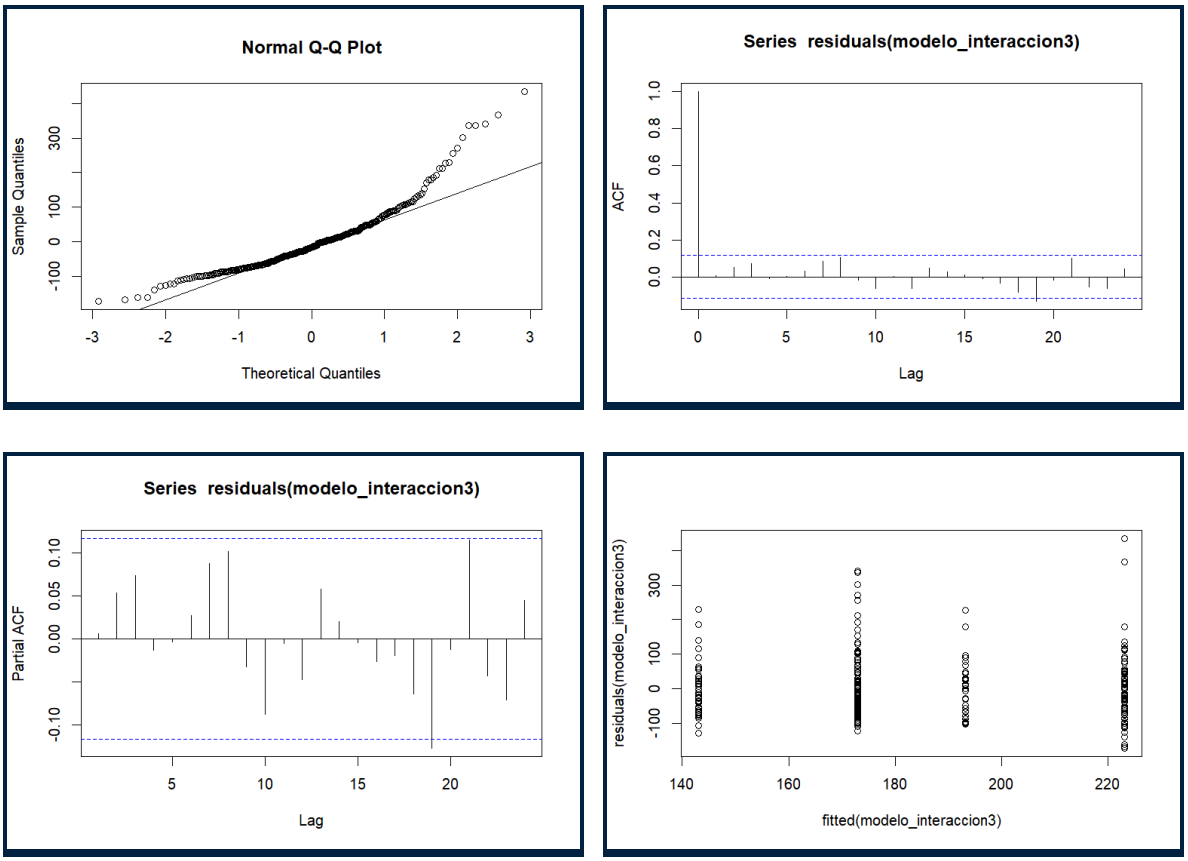
Tras realizar la técnica del Stepwise y AIC para el cortisol en sangre y en pelo, obtenemos dos modelos diferentes con variables significativas para ambos. Por una parte, en el modelo del cortisol en sangre se presentan como variables significativas las siguientes: **AROPE_cat y Smoking**, quedando el modelo como:

$$Cortisol\_sang = \beta 0 + \beta 1 \cdot AROPE\_cat + \beta 2 \cdot Smoking + U$$

Para comprobar que los modelos se consideran ruido blanco, en primer lugar vemos los siguientes gráficos que nos ayudarán a comprobar la validez del modelo del cortisol en sangre:

Tabla 6: *Gráficos de normalidad, autocorrelación y varianza para el modelo del cortisol en sangre*

Fuente: *Elaboración propia*



En esta tabla de gráficos se busca comprobar si el modelo tiene ruido blanco. Para empezar, en el primer gráfico podemos observar que los residuos no están muy bien alineados, por lo tanto se puede deducir que no tiene normalidad. Además, mediante los gráficos ACF y PACF, observamos que no tiene autocorrelación ya que ningún coeficiente de relación sobrepasa el límite de las pruebas de hipótesis, lo que nos demuestra que no tiene autocorrelación. Por otra parte, en el último gráfico se puede observar que la varianza es constante, lo que determina homocedasticidad, y finalmente, tras hacer los cálculos mediante R, se ha comprobado que la media es aproximadamente 0. Por lo tanto, al comprobar que sus valores no son normales podemos decir que el modelo **no es adecuado**.

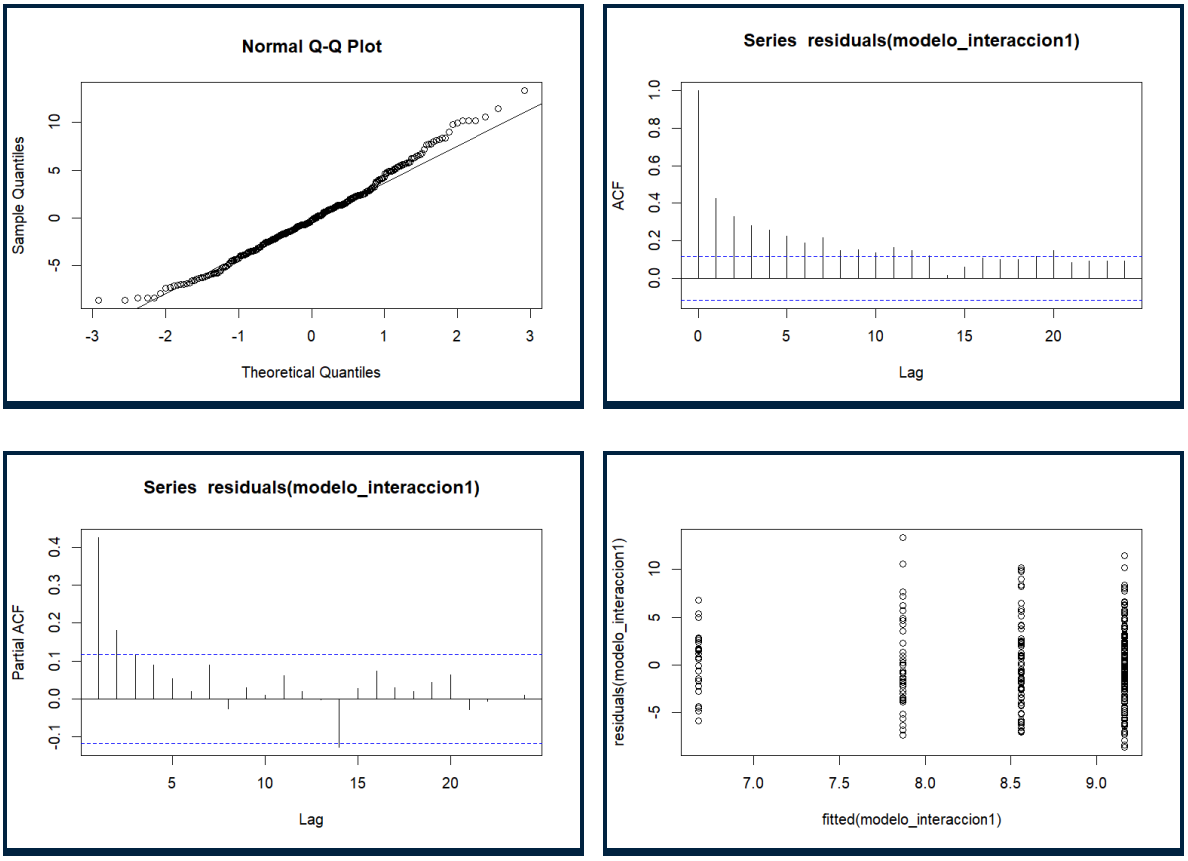
Para el modelo del cortisol en pelo aparecen como significativas las siguientes variables: **M_Cycle, Oral, Drinking, Smoking, partner12 y MicroAlterPos.**, quedando el modelo como:

$$Cortisol\_pelo = \beta 0 + \beta 1 \cdot M\_Cycle + U$$

Por lo tanto, procedemos a comprobar su validez mediante los siguientes gráficos:

Tabla 7: *Gráficos de normalidad, autocorrelación y varianza para el modelo del cortisol en pelo*

Fuente: *Elaboración propia*



En esta tabla de gráficos se busca comprobar si el error del modelo es ruido blanco. Para empezar, en el primer gráfico podemos observar que los residuos no están muy bien alineados. Para comprobar que tiene normalidad se realiza el Kolmogorov-Smirnov normality test, que nos da un resultado del p.value= 0,6098 lo que significa que sigue una distribución normal ya que se acepta la hipótesis nula. Además, mediante los gráficos ACF y PACF, observamos que tiene autocorrelación ya que muchos coeficientes de relación sobrepasan el límite de las pruebas de hipótesis, lo que nos demuestra que tiene autocorrelación. Por otra parte, en el último gráfico se puede observar que la varianza es constante, lo que determina homocedasticidad y finalmente, tras hacer los cálculos mediante R, se ha comprobado que la media es aproximadamente 0. Por todo ello, podemos decir que el modelo **no es adecuado** ya que sus valores tienen autocorrelación.

Tras ver los resultados obtenidos, se considera que transformar la variable dependiente mediante logaritmos es una de las mejores opciones para mejorar la precisión y la fiabilidad de un modelo de regresión. Al aplicar el logaritmo a la variable dependiente, los valores altos de esta se comprimen lo que permite que los efectos extremos tengan un menor impacto y se reduzca la asimetría en los datos.

Además, aunque no se necesite mejorar en estos casos en concreto, utilizar el logaritmo en la variable dependiente es útil para manejar problemas de heterocedasticidad que ocurren cuando la variabilidad de los errores no es constante. Al transformar la variable dependiente se estabiliza la varianza de los residuos, lo que da lugar a estimaciones más consistentes y fiables en diferentes rangos de la variable.

La transformación logarítmica también ayuda a cumplir con las suposiciones estadísticas clave de normalidad y homocedasticidad en el modelo. Esto permite realizar pruebas de significancia de forma más precisa, mejorando la fiabilidad de las inferencias estadísticas. Dado que el modelo se ajusta mejor a una relación lineal cuando la variable dependiente está en escala logarítmica, la regresión resultante refleja de forma más coherente las tendencias subyacentes, optimizando así las predicciones.

Por todo ello, se decide aplicar el logaritmo a la variable dependiente no solo para mejorar el ajuste del modelo, sino también para estabilizar los resultados y aumentar la solidez del análisis estadístico convirtiéndolo en la opción más efectiva en muchos contextos de modelado.

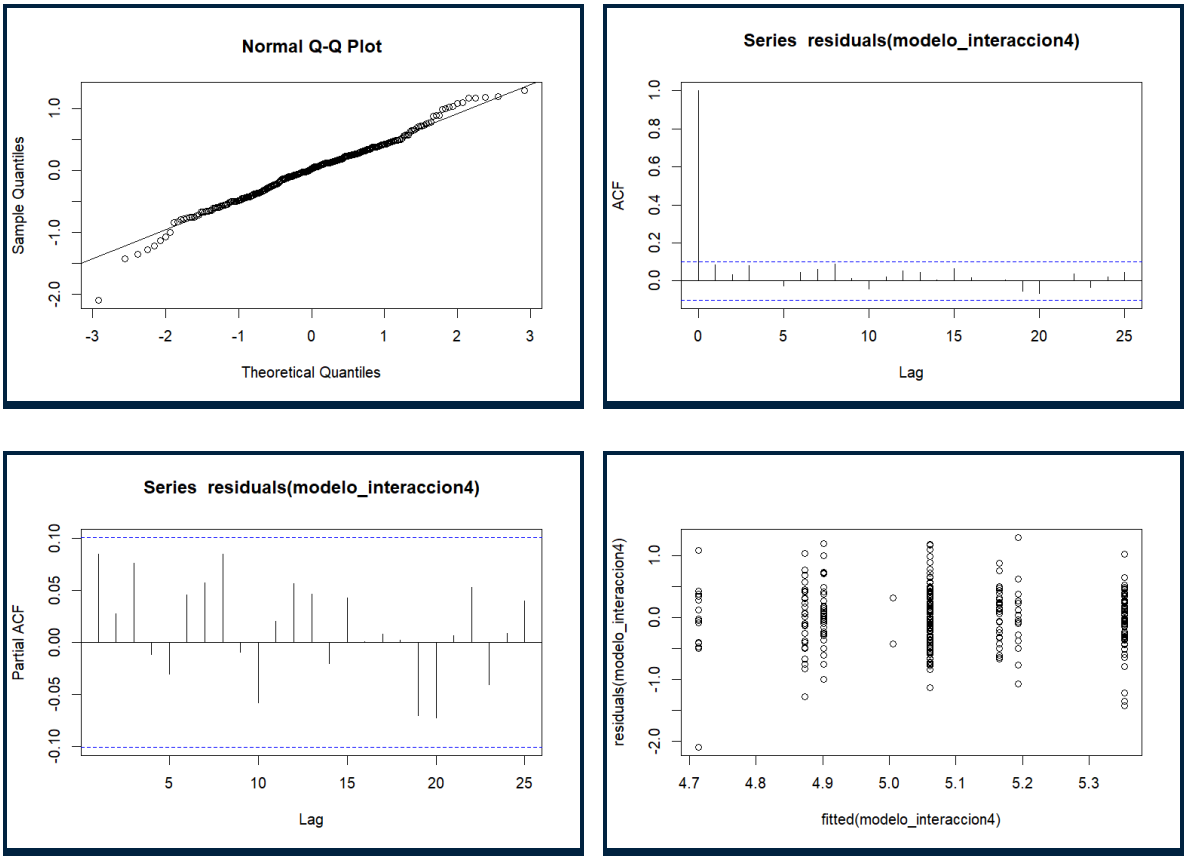
A continuación se va a comprobar si mediante la realización del logaritmo los nuevos modelos mejoran como lo esperado o si por lo contrario, siguen igual o incluso empeoran sus resultados.Para comenzar, veremos si ha mejorado el modelo del cortisol en sangre con el logaritmo. Después de realizar los pasos anteriormente expuestos se llega al modelo final de:

$$Log(Cortisol\_sang) = \beta 0 + \beta 1 \cdot AROPE\_cat + \beta 2 \cdot Smoking + \beta 3 \cdot IUD + U$$

A continuación se comprobará si el modelo del cortisol en sangre con el logaritmo ha mejorado:

Tabla 8: *Gráficos de normalidad, autocorrelación y varianza para el modelo del logaritmo del cortisol en sangre*

Fuente: *Elaboración propia*



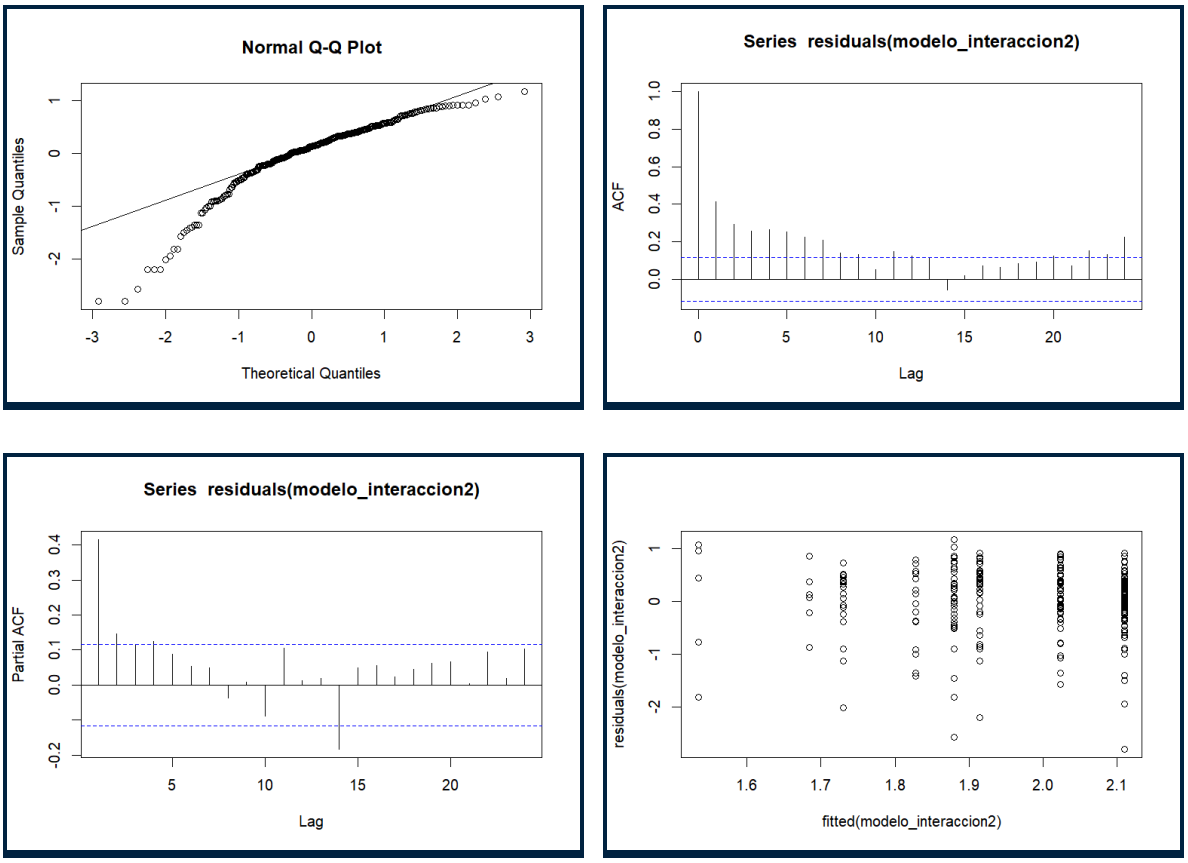
En esta tabla de gráficos se busca comprobar si el error del modelo es ruido blanco. Para empezar, en el primer gráfico podemos observar que los residuos están bastante bien alineados. Para comprobarlo lo hacemos mediante el test de normalidad Kolmogorov-Smirnov con un resultado de un pvalue= 0,6586 por lo tanto podemos aceptar la hipótesis nula y concluir que el modelo sigue una distribución normal. Además, mediante los gráficos ACF y PACF, observamos que no tiene autocorrelación ya que tan solo el primer coeficiente de relación sobrepasa el límite de las pruebas de hipótesis, lo que nos demuestra que no tiene autocorrelación. Por otra parte, en el último gráfico se puede observar que la varianza es constante, lo que determina homocedasticidad, y finalmente, tras hacer los cálculos mediante R se ha comprobado que la media es aproximadamente 0. Por todo ello, podemos decir que el modelo **sí es adecuado**, ya que cumple con la validación del ruido blanco siendo: normal, sin autocorrelación, varianza constante (homocedasticidad) y media aproximadamente 0.

Finalmente, vamos a ver si modificando el modelo con el logaritmo, el modelo de cortisol en pelo ha mejorado así como lo ha hecho el modelo del cortisol en sangre. En este caso, el modelo final quedaría como:

$$Log(Cortisol\_pelo) = \beta 0 + \beta 1 \cdot M\_Cycle + \beta 2 \cdot Smoking + U$$

Tabla 9: *Gráficos de normalidad, autocorrelación y varianza para el modelo del logaritmo del cortisol en pelo*

Fuente: *Elaboración propia*



En esta tabla de gráficos se busca comprobar si el error del modelo es ruido blanco. Para empezar, en el primer gráfico podemos observar que los residuos no están muy bien alineados, por lo tanto se puede deducir que no tiene normalidad. Además, mediante los gráficos ACF y PACF, observamos que tiene autocorrelación ya que muchos de los coeficientes de relación sobrepasan el límite de las pruebas de hipótesis, lo que nos demuestra que tiene autocorrelación. Por otra parte, en el último gráfico se puede observar que la varianza es constante, lo que determina homocedasticidad, y finalmente, tras hacer los cálculos mediante R se ha comprobado que la media es

aproximadamente 0. Por todo ello podemos decir que el modelo **no es adecuado**, ya que sus valores no son normales y tienen autocorrelación.

Tras la validación de todos los modelos se concluye que el único modelo cuyo error es ruido blanco y por lo tanto, el único con el que se puede seguir trabajando, es el modelo del logaritmo del cortisol en sangre quedando como:

$$Log(Cortisol\_sang) = \beta 0 + \beta 1 \cdot AROPE\_cat + \beta 2 \cdot Smoking + \beta 3 \cdot IUD + U$$

4.2 Interpretación del modelo

El siguiente modelo explica los niveles de cortisol en sangre a través de las variables: AROPE_cat, Smoking y IUD. Es decir, vamos a comprobar cómo afecta al nivel de cortisol en sangre de una persona en función de las variables de riesgo de exclusión social o en situación de pobreza, si fuma y si lleva un dispositivo intrauterino (DIU). El modelo se muestra como:

$$log(Cortisol\_sang) = \beta 0 + \beta 1 \cdot AROPE\_cat + \beta 2 \cdot Smoking + \beta 3 \cdot IUD + U$$

En este caso, los modelos con variables dependientes logarítmicas indican cambios porcentuales o proporcionales en lugar de cambios absolutos, por lo tanto se expresarán en forma de porcentajes.

$\beta 0$ Es el logaritmo del promedio de la cantidad de cortisol en sangre cuando una persona no se encuentra en situación de pobreza o exclusión social, no fuman y no utilizan dispositivo intrauterino. Es decir, cuando todas las variables toman valor 0, el logaritmo del nivel promedio del cortisol en sangre es de 5,00392 unidades, que aplicando el exponencial, obtendremos que el nivel de cortisol en sangre es 148,68 nmol/L aproximadamente.

$\beta 1$ indica que el logaritmo del cortisol en sangre de las personas en situación de exclusión social, aumenta en 0,31642 unidades en comparación de quienes no se encuentran en riesgo de exclusión social. En términos porcentuales, los niveles medios de cortisol en sangre de esa persona aumentaría en un 37% (exponencial(0,31642)) con respecto a una persona que no está en exclusión social, siempre y cuando el resto de variables se mantengan constantes.

$\beta 2$ indica que el logaritmo del cortisol en sangre de las personas fumadoras, disminuye en 0,18679 unidades en comparación de quienes no fuman. En términos porcentuales, los niveles medios de cortisol en sangre de las personas fumadoras disminuiría en un 17% (exponencial(0,18679)) con respecto a una persona que no fuma, siempre y cuando se mantengan las demás variables constantes.

β3 indica que el logaritmo del cortisol en sangre de las personas que usan este dispositivo, disminuye en 0,15279 unidades en comparación de quienes no lo usan. En términos porcentuales, los niveles medios de cortisol en sangre de las personas que usan el DIU disminuiría en un 14% (exponencial(0,15279)) con respecto a una persona que no lo utiliza, siempre y cuando se mantengan las demás variables constantes.

Finalmente, para explicar la validez que tiene el modelo del logaritmo del cortisol en sangre se realiza una tabla con los valores más relevantes que se pueden extraer:

***Tabla 10**: Medidas finales del modelo del logaritmo del cortisol en sangre*
***Fuente:** Elaboración propia*

R^2	0,117
R^2 ajustado	0,1075
Raíz del error cuadrático medio	0,4917203
Error absoluto medio	0,3785641
P valor	1,276e-07

Para estos valores hay que tener en cuenta que explican el **logaritmo** del cortisol en sangre. La tabla indica que aproximadamente el 11,7% de la variabilidad del modelo de logaritmo del cortisol en sangre se explica mediante este modelo. Es un valor bajo, lo que podría suponer que no capture toda la variabilidad. Si se observa el R^2 ajustado se aprecia que disminuye ligeramente el valor inicial del R^2 lo que puede suponer que la adición de variables no ha mejorado notoriamente el modelo. Además, la raíz del error cuadrático medio mide la magnitud promedio de los errores de la predicción, siendo en este caso un valor muy bajo lo que implica que las predicciones de modelo tan solo desvían en 0,4917203 unidades de los valores reales. En el caso del error absoluto medio, es el promedio de las desviaciones entre los valores predichos y los obtenidos, y como se indica es menor que la raíz del error cuadrático medio, por lo tanto se puede concluir que los errores están distribuidos de forma razonable. Finalmente, el pvalor de 1,276e-07 nos muestra que el modelo es estadísticamente significativo ya que se encuentra por debajo de 0,05.

4.3 Predicción

En cuanto a la predicción de valores se refiere, una vez construido y validado el modelo de regresión lineal para predecir los niveles de cortisol en sangre, se procede a realizar una nueva base de datos con valores aleatorios para las variables significativas identificadas en el análisis: AROPE_cat, Smoking y IUD. Como en este caso hay 3 variables significativas, significa que mediante 8 valores podemos cubrir todas las posibilidades, por lo tanto probaremos todas las opciones. Los valores utilizados para la predicción fueron seleccionados dentro de los rangos observados en el conjunto de datos original, por ello los valores escogidos para cada uno de ellos variará entre 0 y 1, donde el 0 hace referencia a las respuesta que son “No” y el 1 hace referencia a las respuestas que son “Sí” . Finalmente, la nueva base de datos está compuesta por 8 filas con los siguientes niveles predictivos del logaritmo del cortisol en sangre y su exponencial dependiendo de los factores que cumpla como muestra la Tabla 11:

***Tabla 11**: Base de datos para la predicción de valores*
***Fuente:** Elaboración propia*

ID	Riesgo de exclusión social	Fumadoras	DIU	Logaritmo del cortisol en sangre	Exponencial
1	0	0	0	5.00392	148.68
2	0	0	1	4.90148	134.69
3	0	1	0	4.87329	130.77
4	0	1	1	4.71313	111.37
5	1	0	0	5.35364	211.49
6	1	0	1	5.19349	180.81
7	1	1	0	5.16529	174.15
8	1	1	1	5.00514	148.08

Mediante estos valores podemos observar cómo se comporta el cortisol en sangre en función de los diferentes valores de las variables significativas. Al obtener los valores en forma de logaritmo, hay que pasarlos a su exponencial para saber el valor real del cortisol en sangre. A continuación, ya se puede ver cómo se comportan las

variables y de esta manera nos permite comparar y entender el impacto que tiene fumar, llevar el DIU o encontrarse en situación de riesgo de pobreza.

De la misma forma que se ha visto en la interpretación del modelo, cuando todas las variables son 0, el cortisol en sangre aproximado es de 148,68 nmol/L, y con la tabla anterior podemos determinar los perfiles con el cortisol más alto y más bajo. El caso del perfil con el cortisol más alto es el ID 5, el cual está en riesgo de exclusión y no fuma ni lleva el DIU. Por el contrario, el perfil con los niveles de cortisol en sangre más bajo es el ID 4, el cual fuma y lleva el DIU pero no se encuentra en riesgo de exclusión.

Esto se puede explicar ya que los niveles de cortisol en sangre son un reflejo de la respuesta del organismo al estrés, por lo que estos niveles están directamente relacionados con factores como pueden ser el estilo de vida, el contexto social en el que se encuentre o su salud reproductiva. En cuanto al consumo de tabaco, se obtiene un impacto directo en los niveles de cortisol debido al efecto de la nicotina la cual puede llegar a eliminar la respuesta del eje HPA (hipotálamo-pituitario-adrenal) y de esta forma disminuir la secreción del cortisol en situaciones de estrés (Tafet, 2016). Además, si nos fijamos en el estudio de Orejudo, Camacho y Vega-Michel (2012), se encuentran diferencias entre los participantes de su estudio procedentes de México y de España, viendo cómo los españoles asocian el consumo del tabaco a un momento de relajación lo que explicaría los niveles de cortisol más bajos para los fumadores.

Los dispositivos intrauterinos también pueden influir en los niveles de cortisol. Los métodos anticonceptivos que liberan hormonas como pueden ser los DIUs hormonales interfieren en los niveles de cortisol disminuyendo, de la misma forma que el tabaco, la actividad del eje HPA en respuesta a situaciones estresantes. Por ello se puede explicar que las personas que utilizan este método anticonceptivo hormonal presentan niveles más bajos de cortisol en comparación con otras personas que no lo usan como se puede observar en el estudio de Vázquez et al. (2022).

Sin embargo, en cuanto al riesgo de exclusión social este actúa como un factor estresante que eleva el nivel del cortisol en sangre debido a que se encuentran en estrés constante mayoritariamente por dificultades económicas o por falta de acceso a servicios. Esta situación hace que estas personas se encuentren con una activación constante del eje HPA, lo que genera unos altos niveles de cortisol (Haushofer y Fehr,

2014). Además, como indican varios estudios (Knezevic et al., 2023; Law y Clow, 2020; Cortés, 2006), el estrés debido a razones económicas puede llevar a efectos negativos tanto en la salud mental como en la física, aumentando la predisposición a tener enfermedades relacionadas con niveles altos de cortisol en sangre.

Si analizamos los resultados obtenidos, podemos observar que van cambiando en función de las situaciones en las que se encuentren. Como se ha podido comprobar, el perfil con mayores niveles de cortisol en sangre serían las personas con riesgo de exclusión social, por el estrés constante al que se encuentran. Después, se encontrarán los perfiles con dispositivos intrauterinos hormonales, los cuales pueden modificar la actividad del eje HPA y de esta forma disminuir los niveles de cortisol, y por último se encuentran las personas que consumen tabaco, ya que como se ha expuesto anteriormente, pueden llegar a crear una adaptación y tolerancia de eje HPA disminuyendo también los niveles de cortisol. Además, mediante la tabla también podemos ver cómo los resultados de los niveles de cortisol en sangre se van modificando dependiendo de las características que cumpla, es decir, cómo va aumentando o disminuyendo en función de la relación que tengan las variables con el cortisol. Por ejemplo, si comparamos el caso de ID 5 con el caso del ID 6, podemos ver que si se encuentra en riesgo de exclusión pero lleva el DIU disminuyen los niveles de cortisol, y si este último lo comparamos con el ID 7 vemos que si tiene riesgo de exclusión social y fuma, pero no lleva el DIU, los niveles de cortisol disminuyen aún más que si llevara el DIU.

En conclusión, tras realizar esta predicción podemos concluir que los niveles de cortisol en sangre dependen de factores como el consumo de nicotina, el uso de anticonceptivos hormonales y el contexto social en el que se encuentren, y cada una de estas variables modificará aumentando o disminuyendo los niveles de cortisol, como se ha explicado anteriormente. Es decir, esta perspectiva permite tener una visión de las diferentes combinaciones que el modelo nos proporciona de la regulación de cortisol.

5. Conclusiones

5.1. Objetivos conseguidos

A lo largo de este proyecto, se ha buscado un modelo que prediga los niveles de cortisol en sangre de las mujeres, dependiendo de unas variables independientes. Para ello, se han tratado los datos y se ha explicado cómo se han obtenido, así como el preprocesamiento de los mismos, hasta llegar a obtener una base de datos que permite analizarlos y ver la relación que tienen con la variable respuesta. De esta manera, se cumple con el objetivo de analizar y extraer los factores socio-económicos que pueden influir en los niveles del cortisol.

Tras la limpieza, se han empleado métodos como el Stepwise y AIC para llegar a formar modelos de regresión lineal. Además, una vez los modelos han sido creados se han validado y se ha seguido trabajando con el único modelo que finalmente resulta ser adecuado. Por último, se ha realizado una predicción de los valores del cortisol en sangre dependiendo de los valores que tomen las variables independientes, por lo tanto se ha llegado al objetivo inicial de la construcción de un modelo válido con el que se pueda realizar la predicción de niveles de cortisol.

Además, de este proyecto se pueden extraer las siguientes conclusiones:

- El nivel promedio del cortisol en sangre de las mujeres de la base de datos final es de 157,92 nmol/L lo que se puede considerar un nivel medio.
- La variable que más altera los niveles del cortisol en sangre es la variable AROPE, haciendo que incremente el promedio de estos niveles en 0,31 nmol/L lo que significa un aumento del 37%.
- La predicción nos muestra que si una mujer fuma o lleva el DIU los niveles de cortisol en sangre disminuyen en un 17% para las fumadoras y un 14% para las que llevan el dispositivo intrauterino.
- A pesar de haber tenido inicialmente muchas variables, la mayoría han resultado no aportar información de valor al estudio por lo que se debería buscar otras variables que sí que puedieran serlo.
- Con tan solo las variables de AROPE, Smoking y DIU, se puede explicar el 11.7% de variabilidad del logaritmo de los niveles de cortisol en sangre.
- Los residuos se encuentran muy cerca de 0 demostrando que el modelo no tiene un sesgo sistemático en sus predicciones.

5.2. Lecciones aprendidas

Tras realizar un proyecto de esta envergadura es imprescindible realizar un repaso analizando los puntos que se han ido corrigiendo a lo largo del trabajo y lo que se ha aprendido mediante la realización del mismo.

A lo largo del desarrollo de este trabajo, una de las lecciones más importantes ha sido la de realizar un adecuado procesamiento de datos. Al tratarse de una base de datos con tantas variables hay que ir con precaución comprobando que modificar una variable no ha supuesto que las demás se alteren, y teniendo en cuenta que al combinar muchas bases de datos, puede producirse errores en la casación de los datos. En este caso fue fundamental prestar atención a los detalles durante las transformaciones, lo que supuso un retraso significativo en los avances del proyecto.

Por otra parte, en cuanto a la construcción de los modelos también se probaron diversos métodos para determinar la forma más adecuada de abordar la relación entre las variables. Esto implicó un proceso de pruebas hasta encontrar la estrategia que permitiera extraer un modelo adecuado y válido para seguir trabajando con él.

Finalmente, cabe destacar el notable progreso en los conocimientos sobre el uso de la herramienta estadística R. Tras este proyecto se disponen de unas habilidades excelentes sobre esta herramienta, aprendiendo a utilizar correctamente las librerías, el código y la obtención de gráficos y datos para poder analizarlos. Esto además ha ayudado a poder extraer conclusiones sólidas basadas en evidencias estadísticas.

5.3. Líneas futuras

En el futuro se considera trabajar con nuevas variables que puedan proporcionar información más valiosa para este estudio, ya que como se ha demostrado, tan solo unas pocas han resultado ser significativas de todas las variables iniciales contempladas. Sería interesante comenzar el proceso con variables que sean significativas e ir añadiendo poco a poco otras variables que también lo resulten.

Otro aspecto a mejorar sería la creación de una base de datos única para evitar perder valores por el camino y que venga toda la información del mismo sitio, creando por ejemplo una página web para agilizar la recogida de la información.

6. Bibliografía

1. Adam, H., Khadija, K., & Suzanne, K. (2022). Stepwise Regression: Definition, Uses, Example, and Limitations. Investopedia.
2. Agostinelli, C. (2002). Robust stepwise regression. Journal of Applied Statistics, 29(6), 825-840.
3. Blakemore, E (2024). ¿Qué es el cortisol, para qué sirve y cuál es su importancia?. National Geographic.
4. Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. Wiley Interdisciplinary Reviews: Computational Statistics, 11(3), e1460.
5. Cortés, A. (2006). Inequidad, pobreza y salud. Colombia médica, 37(3), 223-227.
6. Dagnino, J. (2014). Regresión lineal. Revista Chilena de Anestesiología, 43(2).
7. DRESSL, N. L., ETCHEVEST, L. I., FERREIRO, M., & TORRESANI, M. E. (2018). Cortisol como biomarcador de estrés, hambre emocional y estado nutricional. Revista Nutrición Investiga, 3(1).
8. Faresjö, Å., Theodorsson, E., Stomby, A., Quist, H., Jones, M. P., Östgren, C. J., ... & Faresjö, T. (2024). Higher hair cortisol levels associated with previous cardiovascular events and cardiovascular risks in a large cross-sectional population study. BMC Cardiovascular Disorders, 24(1), 536.
9. Granados, R. M. (2016). Modelos de regresión lineal múltiple. Granada, España: Departamento de Economía Aplicada, Universidad de Granada.
10. Haushofer, J., & Fehr, E. (2014). On the psychology of poverty. science, 344(6186), 862-867.
11. Jiménez, M. R., & Aguilá, N. C. (2017). El ciclo menstrual y sus alteraciones. PediatríaIntegral, 304.
12. Karmiola, S., Cuenyaa, L., & Mustaca, A. E. (2019). Comprendo y siento tu dolor: efectos emocionales de viñetas sobre exclusión social en adultos. Acta de investigación psicológica, 9(1), 108-118.
13. Kennedy-Shaffer, L. (2019). Before p< 0.05 to beyond p< 0.05: using history to contextualize p-values and significance testing. The American Statistician, 73(sup1), 82-90.
14. Knezevic, E., Nenic, K., Milanovic, V., & Knezevic, N. N. (2023). The role of cortisol in chronic stress, neurodegenerative diseases, and psychological disorders. Cells, 12(23), 2726.

15. Laguna, C. (2014). Correlación y regresión lineal. Instituto Aragonés de Ciencias de la Salud, 4, 1-18.
16. Law, R., & Clow, A. (2020). Stress, the cortisol awakening response and cognitive function. International review of neurobiology, 150, 187-217.
17. Lechuga, M. L., Luque, O. G., & Martínez, Ú. F. (2015). Crisis y evolución regional del indicador AROPE en España. Rect@: Revista Electrónica de Comunicaciones y Trabajos de ASEPUMA, 16(2), 90-104.
18. Lethielleux, G., & Bertherat, J. (2020). Síndrome de Cushing. EMC-Tratado de Medicina, 24(4), 1-9.
19. Levine, A., Zagoory-Sharon, O., Feldman, R., Lewis, J. G., & Weller, A. (2007). Measuring cortisol in human psychobiological studies. Physiology & behavior, 90(1), 43-53.
20. Marcano, L., & Fermín, W. (2013). Comparación de métodos de detección de datos anómalos multivariantes mediante un estudio de simulación. Saber, 25(2), 193-201.
21. Medina, F., & Galván, M. (2007). Imputación de datos: teoría y práctica. Cepal.
22. NLM, MedlinePlus. Prueba de cortisol. Disponible en: https://medlineplus.gov/spanish/pruebas-de-laboratorio/prueba-de-cortisol/.
23. Organización Mundial de la Salud. Preguntas y respuestas sobre el estrés. Disponible en: https://www.who.int/es/news-room/questions-and-answers/item/stress.
24. Orejudo, S., Camacho, E., & Vega-Michel, C. (2012). Niveles de cortisol salival y tipos de personalidad de Grossarth-Maticek y Eysenck: Un estudio transcultural. Revista de Psicopatología y Psicología Clínica, 17(2).
25. Quiceno, S. H., Bojanini, E. U., Velasquez, J. M. A., Maya, G. C., & Salazar, L. M. (2016). Cortisol: mediciones de laboratorio y aplicación clínica. Medicina & Laboratorio, 22(3), 147-164.
26. Taechakraichana, N., Jaisamrarn, U., Panyakhamlerd, K., Chaikittisilpa, S., & Limpaphayom, K. K. (2002). Climacteric: concept, consequence and care. Journal of the Medical Association of Thailand= Chotmaihet thangphaet, 85, S1-15.
27. Tafet, G. E. (2016). Psiconeuroendocrinología del estrés y la depresión: interacciones entre factores biológicos, psicológicos, genéticos y ambientales. Acta Psiquiátrica y Psicológica de América Latina, 62(3).

28. Torres Jiménez, A. P., & Torres Rincón, J. M. (2018). Climaterio y menopausia. Revista de la Facultad de Medicina (México), 61(2), 51-58.
29. Venables, W. N., & Ripley, B. D. (2013). Modern applied statistics with S-PLUS. Springer Science & Business Media.
30. Vázquez, M. C., Johnson-Ferguson, L., Zimmermann, J., Baumgartner, M. R., Binz, T. M., Beuschlein, F., ... & Quednow, B. B. (2022). Associations of different hormonal contraceptive methods with hair concentrations of cortisol, cortisone, and testosterone in young women. Comprehensive psychoneuroendocrinology, 12, 100161.
31. Zschimmer & Schwarz (2023). ¿Qué es el cortisol y qué tiene que ver con el estrés? Disponible en: https://www.zschimmer-schwarz.es/noticias/que-es-el-cortisol-y-que-tiene-que-ver-con-el-estres/.

7. Anexo I: Relación del trabajo con los Objetivos de Desarrollo Sostenible de la agenda 2030

Tabla 12: *Grado de relación de trabajo con los Objetivos de Desarrollo Sostenible*
Fuente: *Universitat Politècnica de València*

Objetivos de Desarrollo Sostenible	Alto	Medio	Bajo	No procede
ODS 1. **Fin de la pobreza**.	X			
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar**	X			
ODS 4. **Educación de calidad.**				X
ODS 5. **Igualdad de género.**				X
ODS 6. **Agua limpia y saneamiento**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico**.				X
ODS 9. **Industria, innovación e infraestructuras**				X
ODS 10. **Reducción de las desigualdades.**		X		
ODS 11. **Ciudades y comunidades sostenibles.**				X
ODS 12. **Producción y consumo responsables.**				X
ODS 13. **Acción por el clima.**				X
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres**.				X
ODS 16. **Paz, justicia e instituciones sólidas.**				X
ODS 17**. Alianzas para lograr objetivos.**				X

Objetivos de Desarrollo Sostenible

En este anexo se aborda la relación entre el presente TFG y los Objetivos de Desarrollo Sostenible (ODS) propuestos por la Organización de las Naciones Unidas (ONU). Para ello, primero se explicará qué son estos objetivos, así como su origen, y finalmente se analizará la relación que tiene alguno de ellos con el presente trabajo.

Para comenzar, los ODS fueron firmados en 2015 como parte de la Agenda 2030 y establecen un plan de acción global enfocado en promover un desarrollo equilibrado a nivel social, económico y ambiental. Se firmaron un total de 17 objetivos centrándose en desafíos globales como la pobreza, la salud, el cambio climático, la paz y la desigualdad, entre otros.

Uno de los objetivos que podemos relacionar a este TFG es el tercer objetivo de los ODS, la buena salud. A través de este trabajo, se pretende llegar a un modelo estadístico que permita saber qué factores pueden afectar a los niveles del cortisol, relacionado directamente con el estrés. Por ello, este estudio nos ayuda a saber cómo se podría prevenir tener unos niveles de cortisol altos.

Por otra parte, también podemos relacionarlo con el primer ODS, la erradicación de la pobreza, ya que después de realizar el análisis de las variables y del modelo final, se observa que si una mujer se encuentra en riesgo AROPE, tiene un nivel de cortisol más alto que las que no lo tienen. Esto nos demuestra que las personas en situación de exclusión social y de pobreza se encuentran con niveles de cortisol más altos, impactando de forma directa en su salud.

Finalmente, enlazando con el anterior, también se puede relacionar este TFG con el décimo ODS, la reducción de la desigualdad. Se ha demostrado en este estudio que hay una diferencia entre el cortisol en sangre de las mujeres con problemas económicos y las que no los tienen. Por ello, son necesarias políticas que ayuden a mitigar esta desigualdad y de esta forma, mejorar la salud de las mujeres.

8. Anexos II: Código

8.1. Script R Studio

```
install.packages("haven")
library(haven)
install.packages("dplyr")
library(dplyr)
install.packages("ggplot2")
library(ggplot2)
install.packages("GWalkR")
library(GWalkR)
install.packages("MASS")
library(MASS)
install.packages("corrplot")
library(corrplot)
install.packages("mice")
library(mice)
install.packages("lmtest")
library(lmtest)
install.packages("car")
install.packages("GGally")
library(GGally)
library(car)
library(ggplot2)
library(gridExtra)

datos_spss <- read_sav("Coordenadas_emb_15.sav")
datos_spss
datos_spss$idnum %>% summary(.)

VariablesPAPILONG <- read_sav("Variables_PAPILONG.sav")

data <- read_sav("base_Irina.sav")

data <- readRDS("base_Irina.rds")
head(data)
glimpse(data)
str(BDUNI_NA)

#con la base de datos de variables papilong
summary(VariablesPAPILONG$Age)
min(VariablesPAPILONG$Age)
max(VariablesPAPILONG$Age)
mean(VariablesPAPILONG$Age)
median(VariablesPAPILONG$Age)
sd(VariablesPAPILONG$Age)
```

```
mode_value <- get_mode(VariablesPAPILONG$Age)

#he quitado el NA que no tenía valor en la columna Intercourse
sum(is.na(data))
data <- data[!is.na(data$Intercourse), ]
mean(data$Intercourse)
summary(data$Intercourse)

#con esto se saca la cantidad de gente que hay en porcentaje y normal REALIZADO CON CADA VARIABLE CUALITATIVA
frecuencias <- table(BDUNI_complete$AROPE_cat)
porcentajes <- prop.table(frecuencias)*100
frecuencias
porcentajes

datos <- data.frame(Education = factor(c("Primary studies", "Secondary studies", "University studies")))
class(datos$Education)
mode(datos$Education)
hist(data$Education, main ="Histograma de educacion")
summary(BDUNI$Education)

#cambiar variables DE si o no
data$Drinking <- as.numeric(data$Drinking == "Yes")
data$Cpartner <- as.numeric(data$Cpartner == "Yes")
data$Oral <- as.numeric(data$Oral == "Yes")
data$IUD <- as.numeric(data$IUD == "Yes")
data$Preservat <- as.numeric(data$Preservat== "Yes")
data$Antib <- as.numeric(data$Antib == "Yes")
data$Antif_Antiu <- as.numeric(data$Antif_Antiu == "Yes")
data$Prob <- as.numeric(data$Prob == "Yes")
data$MicroAlterNeg <- as.numeric(data$MicroAlterNeg == "Yes")
data$MicroAlterPos <- as.numeric(data$MicroAlterPos == "Yes")

#VOY A CAMBIAR LAS VARIABLES DE SI O NO POR 0 Y 1
#cambiar variables diferentes a si o no
data$Education <- as.numeric(factor(data$Education,
                                    levels = c("Primary studies" , "Secondary studies", "University studies"),
                           labels = c(1,2,3)))

#de esta forma puedo calcular cuanta gente fuma o no de manera exacta
sum(data$Smoking_numerico == 1 & !is.na(data$Smoking_numerico))
sum(data$Smoking_numerico == 0 & !is.na(data$Smoking_numerico))

#relacionar variables con cortisol
#presentar en frecuencias o porcentajes con el comando table
#histogramas boxplots y summarys para variables cuantitativas
```

```
#tablas de frecuencias, tablas de porcentajes y barplots para variables cualitativas
#estudiar relaciones entre ellas a nivel descriptivo

plot(datos_spss$x, datos_spss$y)

#union base de datos INMA y cortisol
data3 <- merge(data, cortisol_sangre_ALL, by.x = "Idnum", by.y = "Id", all.x = TRUE)
data3 <- distinct(data3)

#cambiar nombre variable cortisol
data3 <- rename(data3, Cortisol_sang = `Cortisol nmol/L`)
data3 <- rename(data3, idnum = Idnum)
data3 <- rename(data3, partner12 = Partner12)

#ELIMINAR VARIABLES QUE NO NOS SIRVEN
VariablesPAPILONG <- select(VariablesPAPILONG, -direccion)
VariablesPAPILONG <- select(VariablesPAPILONG, -piso)
VariablesPAPILONG <- select(VariablesPAPILONG, -calle)
VariablesPAPILONG <- select(VariablesPAPILONG, -numero)
VariablesPAPILONG <- select(VariablesPAPILONG, -puerta)
VariablesPAPILONG <- select(VariablesPAPILONG, -municipio)
VariablesPAPILONG <- select(VariablesPAPILONG, -CP)
VariablesPAPILONG <- VariablesPAPILONG[, -30]
data3 <- select(data3, -Caja)

#Añadir una nueva columna en data3 a 0 y en variablespapilong a 1 para diferenciar de que
bbdd son
data3$BBDD <- 0
VariablesPAPILONG$BBDD <- 1

#para añadir cortisol en pelo a la base de datos ya que no habiamos contado con ella
BDUNI <- BDUNI %>%
  left_join(VariablesPAPILONG %>% select(idnum, Cortisol_pelo), by = "idnum")

BDUNI <- BDUNI %>%
  left_join(data3 %>% select(idnum, Cortisol_pelo), by = "idnum")

library(dplyr)

data3 <- data3 %>%
        left_join(Resultados_Muestras_cabello_Proyecto_INMA    %>%    dplyr::select(idnum,
Cortisol_pelo), by = "idnum")

#cambio de nombre en la base de datos
names(Resultados_Muestras_cabello_Proyecto_INMA)[names(Resultados_Muestras_cabel
lo_Proyecto_INMA) == "Cortisol nmol/L"] <- "Cortisol_pelo"
names(Resultados_Muestras_cabello_Proyecto_INMA)[names(Resultados_Muestras_cabel
lo_Proyecto_INMA) == "Idnum"] <- "idnum"
```

```
Resultados_Muestras_cabello_Proyecto_INMA$idnum <-
as.numeric(Resultados_Muestras_cabello_Proyecto_INMA$idnum)
Resultados_Muestras_cabello_Proyecto_INMA<-
Resultados_Muestras_cabello_Proyecto_INMA[, -2]

#unir las bases de datos
BDUNI <- merge(x = data3, y = VariablesPAPILONG, by = "idnum", all=TRUE)

#unir las variables en una, asi es de una en una
BDUNI$Education <- ifelse(is.na(BDUNI$Education.x), BDUNI$Education.y,
BDUNI$Education.x)
BDUNI$Age <- ifelse(is.na(BDUNI$Age.x), BDUNI$Age.y, BDUNI$Age.x)

#para cambiar todas las variables de una
for (col_name in names(BDUNI)) {
  # Verificamos si el nombre de la columna termina en .x
  if (endsWith(col_name, ".x")) {
    # Extraemos el nombre de la columna original eliminando el sufijo .x
    base_col_name <- substr(col_name, 1, nchar(col_name) - 2)

    # Creamos una nueva columna que combina los valores de .x y .y
    BDUNI[[base_col_name]] <- ifelse(is.na(BDUNI[[paste0(base_col_name, ".x")]]),
                                     BDUNI[[paste0(base_col_name, ".y")]],
                                     BDUNI[[paste0(base_col_name, ".x")]])

    # Eliminamos las columnas .x y .y del data frame
          BDUNI <- BDUNI[, !(names(BDUNI) %in% c(paste0(base_col_name, ".x"),
paste0(base_col_name, ".y")))]
  }
}

#cambiar las variables de health_stat, education y BMI donde tenga un 1 a 0 etc
BDUNI <- BDUNI %>%
  mutate(Health_Stat = ifelse(BBDD == 1,
                              ifelse(Health_Stat == 1, 0,
                                     ifelse(Health_Stat == 2, 1,
                                            ifelse(Health_Stat == 3, 2, Health_Stat))),
                              Health_Stat),
         Education = ifelse(BBDD == 1,
                            ifelse(Education == 1, 0,
                                   ifelse(Education == 2, 1,
                                          ifelse(Education == 3, 2, Education))),
                            Education),
         BMI = ifelse(BBDD == 1,
                      ifelse(BMI == 1, 0,
                             ifelse(BMI == 2, 1,
                                    ifelse(BMI == 3, 2, BMI))),
                      BMI)
```

```
 )

# Convertir la columna a numérica
BDUNI$Cortisol_pelo <- as.numeric(BDUNI$Cortisol_pelo)

#quitamos sc, y workout ya que no hay para las dos bases de datos. También he quitado observations
#caja y codigo de laboratorio tb porque no nos da infor
BDUNI <- select(BDUNI, -SC)
BDUNI <- select(BDUNI, -Workout_perciv_3c)
BDUNI <- select(BDUNI, -observations)
BDUNI <- select(BDUNI, -Caja)
BDUNI <- BDUNI[, -2]

#para preprocesamiento de datos voy a ver cuantos NA hay por columna
na_por_colimna <- colSums(is.na(BDUNI_NA_FACTOR))
print(na_por_colimna)
#PARA VER LOS PORCENTAJES DE NA
PORCENTAJE  <- function(x) {sum(is.na(x)) / length(x)*100}
apply(BDUNI, 2, PORCENTAJE)

#cantidad de valores que hay por columna
BDUNI %>% summarise_all(funs(n_distinct(.)))
unique(BDUNI$Intercourse)

BDUNI_NA <- BDUNI

# Reemplazar los valores "<1,5" por "0.75"
BDUNI$Cortisol_pelo[BDUNI$Cortisol_pelo == "<1,50"] <- "0.75"

BDUNI_NA$Cortisol_pelo[BDUNI_NA$Cortisol_pelo == "<1,50"] <- "0.75"

#quitar filas con valor na en cortisol
data3 <- data3[!is.na(data3$Cortisol_sang), ]
BDUNI <- BDUNI[!is.na(BDUNI$Cortisol_sang), ]
BDUNI <- BDUNI[!is.na(BDUNI$Cortisol_pelo), ]

#para añadir FECHA a la base de datos ya que no habiamos contado con ella
BDUNI <- BDUNI %>%
  left_join(VariablesPAPILONG %>% select(idnum, Drinking), by = "idnum")

BDUNI <- BDUNI %>%
  left_join(data3 %>% select(idnum, Drinking), by = "idnum")

BDUNI$Drinking<-NULL
```

```
#IMPUTACIÓN DE LAS VARIABLES

# Imputar los valores faltantes
imputed_data <- mice(BDUNI, method = "pmm", m = 5, maxit = 50, seed = 500)
# Obtener el dataset completo imputado
BDUNI_complete <- complete(imputed_data)
# Verificar patrones de datos faltantes
md.pattern(BDUNI_complete)
#base de datos sin NAs
BDUNI_NA <- na.omit(BDUNI)

na_por_colimna <- colSums(is.na(BDUNI_NA))
print(na_por_colimna)

#para preprocesamiento de datos voy a ver cuantos NA hay por columna
na_por_colimna <- colSums(is.na(BDUNI_complete))
print(na_por_colimna)

BDUNI_NA$Cortisol_pelo <- round(BDUNI_NA$Cortisol_pelo, 2)

boxplot(BDUNI_NA$Cortisol_pelo ~ BDUNI_NA$BBDD, ylab = "Cortisol en pelo", xlab= "BBDD")
model <- lm(BDUNI_NA$Cortisol_pelo ~ BDUNI_NA$BBDD)
summary(model)
boxplot(BDUNI_NA$Cortisol_sang ~ BDUNI_NA$BBDD, ylab = "Cortisol en sangre", xlab= "BBDD")
model <- lm(BDUNI_NA$Cortisol_sang ~ BDUNI_NA$BBDD)
summary(model)

BDUNI_NA$Cortisol_pelo <- as.numeric(BDUNI_NA$Cortisol_pelo)

summary(BDUNI_NA_LOG_SANG$Age)
summary(BDUNI_NA_LOG_SANG$Intercourse)
boxplot(BDUNI_NA_LOG_SANG$Age)
qqnorm(BDUNI_NA_LOG_SANG$Age)
qqline(BDUNI_NA_LOG_SANG$ge)
boxplot(BDUNI_NA_LOG_SANG$Intercourse)
table(BDUNI_NA_FACTOR$AROPE_cat)

#regresión lineal
model <- lm(BDUNI_NA_FACTOR$Cortisol_pelo ~ BDUNI_NA_FACTOR$Education)
summary(model)
boxplot(BDUNI_NA_FACTOR$Cortisol_pelo ~ BDUNI_NA_FACTOR$Education, ylab = "Cortisol en pelo", xlab= "Educación")
model <- lm(BDUNI_NA_FACTOR$Cortisol_sang ~ BDUNI_NA_FACTOR$Education)
summary(model)
boxplot(BDUNI_NA_FACTOR$Cortisol_sang ~ BDUNI_NA_FACTOR$Education, ylab = "Cortisol en sangre", xlab= "Educación")
```

```
#y asi con todas las demas variables

summary(BDUNI_complete$Intercourse)

hist(BDUNI_NA_FACTOR$Cortisol_sang,  ylab = "Frecuencia", xlab= "Cortisol en sangre",
main="Histograma del Cortisol en sangre")
hist(BDUNI_NA_FACTOR$Cortisol_pelo, ylab = "Frecuencia" , xlab= "Cortisol en pelo",
main="Histograma del Cortisol en pelo")

ggplot(BDUNI_NA_FACTOR, aes( x = Cortisol_sang, y = Cortisol_pelo, color = Smoking))+
      geom_point(alpha = 0.6, size = 3) +  # Puntos de dispersión
        geom_smooth(method = "lm", se = FALSE) +  # Línea de regresión
        labs(title = "Relación entre cortisol en sangre y pelo",
             x = "Cortisol en sangre (nmol/L)",
             y = "Cortisol en pelo (nmol/L)",
             color= "Fumadores") +
  theme_minimal()

ggplot(BDUNI_NA_FACTOR, aes(x = Smoking, y = Cortisol_sang, fill = Smoking)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Distribución del cortisol en sangre por fumadores",
       x = "Smoking",
       y = "Cortisol en sangre (nmol/L)") +
  theme_minimal()

#t test y ANOVA

#A CONTINUACION TODO JUNTO
#gráfico de correlaciones de todas las variables con CORTISOL
cor_matrix <- cor(BDUNI, use = "complete.obs")
cortisol_cor <- cor_matrix[, "Cortisol_sang"]
# Convertir a data frame para ggplot2
cortisol_cor_df <- as.data.frame(cortisol_cor)
cortisol_cor_df$variable <- rownames(cortisol_cor_df)
# Eliminar la fila de la variable Cortisol_sang
cortisol_cor_df <- cortisol_cor_df[cortisol_cor_df$variable != "Cortisol_sang", ]
# Crear gráfico de barras
ggplot(cortisol_cor_df, aes(x = reorder(variable, Cortisol_sang), y = Cortisol_sang)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  coord_flip() +
   labs(title = "Correlaciones con la variable Cortisol_sang", x = "Variables", y = "Correlación")
+
  theme_minimal()
# Gráfico de correlaciones con corrplot
corrplot(cor_matrix, method = "circle", type = "upper", tl.col = "black", tl.srt = 45, addCoef.col
= "black",
         tl.cex = 0.3, # Tamaño de las etiquetas
         mar = c(2, 2, 2, 2), # Márgenes externos
```

```
        title = "Gráfico de correlaciones", # Título del gráfico
        tl.offset = 2, # Ajuste de la posición de las etiquetas
        number.cex = 0.3) # Tamaño de los coeficientes de correlación)

# Seleccionar las columnas que se quieren incluir en el análisis, excluyendo algunas
variables_incluidas <- BDUNI_complete[, !names(BDUNI_complete) %in% c("V_Date", "M_Date")]
# Asegurarse de que solo se incluyan las columnas numéricas
numeric_columns <- sapply(variables_incluidas, is.numeric)
variables_numericas <- variables_incluidas[, numeric_columns]
# Calcular la matriz de correlación
cor_matrix <- cor(variables_numericas, use = "complete.obs")
# Extraer las correlaciones con la variable CORTISOL
cortisol_cor <- cor_matrix[, "Cortisol_pelo", drop = FALSE]
# Graficar las correlaciones con CORTISOL
corrplot(cor(cortisol_cor), method = "circle", type = "upper", tl.col = "black", tl.srt = 45, addCoef.col = "black")

ggplot(cortisol_cor, aes(x = reorder(Variable, Correlation), y = Correlation)) +
  geom_bar(stat = "identity", fill = "black") +
  coord_flip() +
  labs(title = "Correlaciones de las variables con CORTISOL",
       x = "Variable",
       y = "Correlación") +
  theme_minimal()

# Calcular la matriz de correlación entre Cortisol_sang y todas las demás variables
cor_matrix <- cor(BDUNI[, !names(BDUNI) %in% "Cortisol_sang"], BDUNI$Cortisol_sang)
# Cargar la librería corrplot si no está cargada
if (!requireNamespace("corrplot", quietly = TRUE)) {
  install.packages("corrplot")
}
library(corrplot)
# Gráfico de correlaciones con corrplot
corrplot(cor_matrix, method = "circle", type = "upper", tl.col = "black", tl.srt = 45, addCoef.col = "black") # Tamaño de los coeficientes de correlación
any(is.na(cor_matrix))


ggplot(cortisol_cor, aes(x = variable, y = correlation, fill = variable)) +
  geom_bar(stat = "identity") +  # o geom_point() para un gráfico de puntos
  labs(title = "Correlación con Cortisol_pelo",
       x = "Variable",
       y = "Correlación") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

hist(BDUNI_complete$Cortisol_pelo)
```

```
hist(BDUNI_complete$Cortisol_sang)
hist(log(BDUNI_complete$Cortisol_pelo))
hist(log(BDUNI_complete$Cortisol_sang))

#scatter plot, grafico correlacion de todas
variables <- colnames(BDUNI_complete)
variables <- variables[variables != "Cortisol_pelo"]  # Excluir la variable Cortisol_sang

#stepwise

lm1 <- lm(Cortisol_sang ~., data = BDUNI_complete )
slm1 <- step(lm1)

lm2 <- lm(Cortisol_sang ~1 , data = BDUNI_complete )
slm2 <- step(lm2,  scope = lm1, direction = c("forward"))

#copia de la bbdd sin las variables que no sean significativas
BDUNI_final <- BDUNI_complete
BDUNI_final <- BDUNI_final[, -1]
BDUNI_NA <- BDUNI_NA[, -24]

BDUNI_NA_LOG_SANG <- BDUNI_NA_LOG_SANG[, -1]

BDUNI_NA_FACTOR <- NULL

# Nombres de las variables que no deseas convertir a factor
excluir <- c("Age", "Intercourse", "AROPE_cont", "Cortisol_pelo")
# Crear una copia del data frame para evitar modificar el original
BDUNI_final_factor <- BDUNI_final
BDUNI_NA_FACTOR <- BDUNI_NA

BDUNI_NA_PELO <- BDUNI_NA_PELO %>%
  mutate(across(-all_of(excluir), as.factor))

BDUNI_NA_PELO$Cortisol_pelo <- as.numeric(BDUNI_NA_PELO$Cortisol_pelo)

# Convertir todas las columnas a factores, excepto las excluidas
BDUNI_NA[] <- lapply(BDUNI_NA, function(x) {
  if (is.character(x) && !(names(BDUNI_NA) %in% excluir)) {
    as.factor(x)
  } else {
    x
  }
})

BDUNI_NA_PELO$Cortisol_pelo <- as.numeric(BDUNI_NA_PELO$Cortisol_pelo)

# Convertir las variables a factores excepto las especificadas
```

```
BDUNI_NA <- as.data.frame(lapply(names(BDUNI_NA), function(col) {
  if (col %in% excluir) {
    return(BDUNI_NA[[col]])
  } else {
    return(as.factor(BDUNI_NA[[col]]))
  }
}))

# Restaurar los nombres de las columnas
names(BDUNI_NA_FACTOR) <- names(BDUNI_NA)

# Mostrar un resumen para verificar
str(BDUNI_NA_PELO)

BDUNI_NA_FACTOR <- BDUNI_NA

# Crear la base de datos BDUNI_final_sang sin la variable Cortisol_pelo
BDUNI_NA_SANG <- BDUNI_NA[, !names(BDUNI_NA) %in% "Cortisol_pelo"]

# Crear la base de datos BDUNI_final_pelo sin la variable Cortisol_sang
BDUNI_NA_PELO <- BDUNI_NA[, !names(BDUNI_NA) %in% "Cortisol_sang"]

#base de datos con logaritmos
BDUNI_NA_LOG <- BDUNI_NA_FACTOR

# Transformar las variables Cortisol_pelo y Cortisol_sang a sus logaritmos
BDUNI_NA_LOG$Cortisol_pelo <- log(BDUNI_NA_LOG$Cortisol_pelo)
BDUNI_NA_LOG$Cortisol_sang <- log(BDUNI_NA_LOG$Cortisol_sang)

BDUNI_NA_LOG_PELO <- BDUNI_NA_LOG
BDUNI_NA_LOG_PELO$Cortisol_sang <- NULL  # Eliminar la columna Cortisol_sang

# Crear la base de datos con Cortisol_sang y todas las demás variables
BDUNI_NA_LOG_SANG <- BDUNI_NA_LOG
BDUNI_NA_LOG_SANG$Cortisol_pelo <- NULL  # Eliminar la columna Cortisol_pelo

# Verificar los cambios
head(BDUNI_log_pelo)

# Convertir variables cualitativas a factores si es necesario
# df$variable_cualitativa <- as.factor(df$variable_cualitativa)

# Identificar variables cuantitativas y cualitativas
quantitative_vars <- names(BDUNI_final_sang)[sapply(BDUNI_final_sang, is.numeric)]
qualitative_vars <- names(BDUNI_final_sang)[sapply(BDUNI_final_sang, is.factor)]

quantitative_vars <- names(BDUNI_final_pelo)[sapply(BDUNI_final_pelo, is.numeric)]
qualitative_vars <- names(BDUNI_final_pelo)[sapply(BDUNI_final_pelo, is.factor)]
```

```
quantitative_vars <- names(BDUNI_log_pelo)[sapply(BDUNI_log_pelo, is.numeric)]
qualitative_vars <- names(BDUNI_log_pelo)[sapply(BDUNI_log_pelo, is.factor)]

quantitative_vars <- names(BDUNI_log_sang)[sapply(BDUNI_log_sang, is.numeric)]
qualitative_vars <- names(BDUNI_log_sang)[sapply(BDUNI_log_sang, is.factor)]

df_quantitative <- BDUNI_log_sang[, quantitative_vars]
df_qualitative <- BDUNI_final_pelo[, qualitative_vars]
ggpairs(df_quantitative,
        lower = list(continuous = wrap("smooth", method = "lm", se = FALSE)),
        upper = list(continuous = "cor"),
        diag = list(continuous = "densityDiag")) +
  theme_minimal()

df_quantitative <- df_quantitative[!is.na(df_quantitative$Cortisol_sang), ]

# Eliminar la variable 'Cortisol_pelo' de las listas
quantitative_vars <- setdiff(quantitative_vars, "Cortisol_pelo")
qualitative_vars <- setdiff(qualitative_vars, "Cortisol_pelo")

# Crear gráficos de dispersión para variables cuantitativas
scatter_plots <- lapply(BDUNI_final, function(var) {
  ggplot(BDUNI_final, aes_string(x = var, y = "Cortisol_pelo")) +
    geom_point() +
    geom_smooth(method = "lm", col = "red") +
    labs(title = paste("Scatter plot of Cortisol_pelo vs", var),
         x = var, y = "Cortisol_pelo") +
    theme_minimal()
})

# Crear gráficos de cajas para variables cualitativas
box_plots <- lapply(BDUNI_final, function(var) {
  ggplot(BDUNI_final, aes_string(x = var, y = "Cortisol_pelo")) +
    geom_boxplot() +
    labs(title = paste("Box plot of Cortisol_pelo vs", var),
         x = var, y = "Cortisol_pelo") +
    theme_minimal()
})

num_cols <- 6 # Número de columnas para los gráficos
plot_list <- c(scatter_plots, box_plots)

# Función para añadir márgenes alrededor de los gráficos
add_margins <- function(plot) {
  plot + theme(plot.margin = unit(c(1, 1, 1, 1), "cm"))
}
```

```
# Aplicar márgenes a cada gráfico
plot_list <- lapply(plot_list, add_margins)

# Mostrar los gráficos
grid.arrange(grobs = plot_list, ncol = num_cols)

par(mar = c(1, 1, 1, 1))
pairs(df_quantitative)

# Filtrar las primeras variables para dividir la matriz
first_vars <- names(df_qualitative)[1:5]  # Elegir las primeras 6 variables
second_vars <- names(df_qualitative)[6:10]
third_vars <- names(df_qualitative)[11:17]
fourth_vars <- names(df_qualitative)[18:22]

df_quantitative <- na.omit(df_quantitative)

pairs(df_quantitative,
      upper.panel = panel.reg,
      diag.panel = panel.hist,
      lower.panel = panel.cor,
      main = "Matriz de Scatterplots",
      cex.main = 1.2)

head(BDUNI_final$Cortisol_pelo)
unique(BDUNI_final$Cortisol_pelo)

str(BDUNI_NA_SANG)

#STEPWISE
# Cargar el paquete necesario
library(MASS)

# Crear el modelo base con solo el intercepto
modelo_base <- lm((Cortisol_sang) ~ 1, data = BDUNI_NA_SANG)

# Crear el modelo completo con todas las variables explicativas
modelo_completo <- lm((Cortisol_sang) ~ ., data = BDUNI_NA_SANG)

# Selección de variables hacia adelante
modelo_step_forward <- stepAIC(modelo_base,
                               scope = list(lower = modelo_base, upper = modelo_completo),
                               direction = "forward")

# Selección de variables hacia atrás
modelo_step_backward <- stepAIC(modelo_completo,
                                scope = list(lower = modelo_base, upper = modelo_completo),
                                direction = "backward")
```

```
# Selección de variables en ambas direcciones
modelo_step_both <- stepAIC(modelo_base,
                    scope = list(lower = modelo_base, upper =  modelo_completo),
                    direction = "both")

# Imprimir los resultados de cada modelo
summary(modelo_step_forward)
summary(modelo_step_backward)
summary(modelo_step_both)

#ahora con cortisol en pelo
modelo_pelo <- lm((Cortisol_pelo) ~ 1, data = BDUNI_NA_PELO)

# Crear el modelo completo con todas las variables explicativas
modelo_completo_pelo <- lm((Cortisol_pelo) ~ ., data = BDUNI_NA_PELO)

# Selección de variables hacia adelante
modelo_step_forward_pelo <- stepAIC(modelo_pelo,
                     scope = list(lower = modelo_pelo, upper = modelo_completo_pelo),
                     direction = "forward")

# Selección de variables hacia atrás
modelo_step_backward_pelo <- stepAIC(modelo_completo_pelo,
                      scope = list(lower = modelo_pelo, upper = modelo_completo_pelo),
                      direction = "backward")


# Selección de variables en ambas direcciones
modelo_step_both_pelo <- stepAIC(modelo_pelo,
                    scope = list(lower = modelo_pelo, upper = modelo_completo_pelo),
                    direction = "both")

# Imprimir los resultados de cada modelo
summary(modelo_step_forward_pelo)
summary(modelo_step_backward_pelo)
summary(modelo_step_both_pelo)

#modelo 6
modelo_interaccion6 <- lm(Cortisol_pelo ~ (M_Cycle + Oral + Drinking + Smoking +
partner12 + MicroAlterPos)^2, data = BDUNI_NA_PELO)
summary(modelo_interaccion6)

##PARA VER SI HAY CORRELACIONES ENTRE LAS VARIABLES
install.packages("olsrr")
library(olsrr)
```

```
model_full <- lm(Cortisol_pelo ~ (M_Cycle + Oral + Drinking + Smoking + partner12 + MicroAlterPos)^2,
                 BDUNI_NA_PELO)

models_both <- ols_step_both_p(model_full,
                               details = TRUE,
                               p_enter = 0.06,
                               p_remove = 0.1)

#MODELO FINAL
modelo_interaccion1 <- lm(Cortisol_pelo ~ M_Cycle , data = BDUNI_NA_PELO)
summary(modelo_interaccion1)
anova(modelo_step_both_pelo, modelo_interaccion1)

#con 0.1
Cortisol_pelo ~ M_Cycle + Oral + Smoking:partner12 + partner12 + M_Cycle:partner12 + M_Cycle:MicroAlterPos

#interaccion entre las variables CORTISOL EN PELO CON LOGS SIN NA

model_full <- lm(Cortisol_pelo ~ (M_Cycle + Smoking + Oral + Drinking )^2,
                 BDUNI_NA_LOG_PELO)

models_both <- ols_step_both_p(model_full,
                               details = TRUE,
                               p_enter = 0.06,
                               p_remove = 0.1)

#MODELO FINAL

modelo_interaccion2 <- lm(Cortisol_pelo ~ Smoking + M_Cycle , data = BDUNI_NA_LOG_PELO)

summary(modelo_interaccion2)

#con 0.1
modelo_interaccion2 <- lm(Cortisol_pelo ~ Smoking + M_Cycle + Oral + Drinking , data = BDUNI_NA_LOG_PELO)
summary(modelo_interaccion2)

#interaccion entre las variables CORTISOL EN SANGRE SIN LOGS SIN NA
model_full <- lm(Cortisol_sang ~ (AROPE_cat + Oral + Smoking )^2,
                 BDUNI_NA_SANG)

models_both <- ols_step_both_p(model_full,
                               details = TRUE,
                               p_enter = 0.06,
                               p_remove = 0.1)
```

```
#MODELO FINAL
modelo_interaccion3 <- lm(Cortisol_sang ~ Smoking + AROPE_cat, data = BDUNI_NA_SANG)
summary(modelo_interaccion3)

#interaccion entre las variables CORTISOL EN SANGRE CON LOGS SIN NA
model_full <- lm(Cortisol_sang ~ (AROPE_cat + Oral + Smoking + IUD + Health_Stat)^2,
                 BDUNI_NA_LOG_SANG)

models_both <- ols_step_both_p(model_full,
                               details = TRUE,
                               p_enter = 0.06,
                               p_remove = 0.1)

#MODELO FINAL
modelo_interaccion4 <- lm(Cortisol_sang ~ AROPE_cat + Smoking + IUD + Health_Stat , data = BDUNI_NA_LOG_SANG)
summary(modelo_interaccion4)
summary(models_both)
plot(models_both)

#VALIDACION CORRECTA

# Cargar los paquetes
library(tseries)
#normalidad
# Kolmogorov-Smirnov Test para modelo_interaccion1
cat("Modelo 1: Kolmogorov-Smirnov Test\n")
print(ks.test(resid_1, "pnorm", mean = mean(resid_1), sd = sd(resid_1)))

# Kolmogorov-Smirnov Test para modelo_interaccion2
cat("Modelo 2: Kolmogorov-Smirnov Test\n")
print(ks.test(resid_2, "pnorm", mean = mean(resid_2), sd = sd(resid_2)))

# Kolmogorov-Smirnov Test para modelo_interaccion3
cat("Modelo 3: Kolmogorov-Smirnov Test\n")
print(ks.test(resid_3, "pnorm", mean = mean(resid_3), sd = sd(resid_3)))

# Kolmogorov-Smirnov Test para modelo_interaccion4
cat("Modelo 4: Kolmogorov-Smirnov Test\n")
print(ks.test(resid_5, "pnorm", mean = mean(resid_5), sd = sd(resid_5)))

# Test de homocedasticidad (Breusch-Pagan) hipotesis nula varianza constante, hipotesis no nula varianza no constante
install.packages("lmtest")
library(lmtest)
bptest(modelo_interaccion3)
```

```
# Autocorrelación: dibujar ACF y PACF de los residuos
acf(residuals(modelo_interaccion1))  # ACF: autocorrelation function
pacf(residuals(modelo_interaccion1)) # PACF: partial autocorrelation function

# Verificar si la media de los residuos es igual a 0 (media cercana a 0)
mean_residuos <- mean(residuals(modelo_interaccion3))
mean_residuos  # Debería estar cerca de 0

# Graficar histograma y boxplot de los residuos
hist(residuals(modelo_interaccion3), main = "Histograma de los residuos", xlab = "Residuos", col = "lightblue", breaks = 20)
boxplot(residuals(modelo_interaccion3), main = "Boxplot de los residuos", horizontal = TRUE, col = "lightblue")
#sin logs la normalidad mal, con logs debería arreglarse. Los demás debería mejorar o quedarse igual

# Modelo 1: modelo_interaccion1
modelo_interaccion1 <- lm(Cortisol_pelo ~  M_Cycle , data = BDUNI_NA_PELO)

# Verificación de los supuestos para modelo_interaccion1
#hist(residuals(modelo_interaccion1))  # Histograma de los residuos
qqnorm(residuals(modelo_interaccion1))  # Q-Q plot de los residuos
qqline(residuals(modelo_interaccion1))  # Línea de referencia para normalidad
shapiro.test(residuals(modelo_interaccion1))  # Prueba de Shapiro-Wilk

# Autocorrelación: dibujar ACF y PACF de los residuos
acf(residuals(modelo_interaccion1))  # ACF: autocorrelation function
pacf(residuals(modelo_interaccion1)) # PACF: partial autocorrelation function
durbinWatsonTest(modelo_interaccion1)  # Prueba de autocorrelación Durbin-Watson

plot(fitted(modelo_interaccion1), residuals(modelo_interaccion1))  # Gráfico de residuos vs valores ajustados
bptest(modelo_interaccion1)# Prueba de homocedasticidad Breusch-Pagan

mean(residuals(modelo_interaccion1))  # Media de los residuos

# Modelo 2: modelo_interaccion2
modelo_interaccion2 <- lm(Cortisol_pelo ~ Smoking + M_Cycle, data = BDUNI_NA_LOG_PELO)

# Verificación de los supuestos para modelo_interaccion2
#hist(residuals(modelo_interaccion2))  # Histograma de los residuos
qqnorm(residuals(modelo_interaccion2))  # Q-Q plot de los residuos
qqline(residuals(modelo_interaccion2))  # Línea de referencia para normalidad
shapiro.test(residuals(modelo_interaccion2))  # Prueba de Shapiro-Wilk

acf(residuals(modelo_interaccion2))  # ACF: autocorrelation function
```

```
pacf(residuals(modelo_interaccion2)) # PACF: partial autocorrelation function
durbinWatsonTest(modelo_interaccion2)  # Prueba de autocorrelación Durbin-Watson

plot(fitted(modelo_interaccion2), residuals(modelo_interaccion2))  # Gráfico de residuos vs valores ajustados
bptest(modelo_interaccion2)  # Prueba de homocedasticidad Breusch-Pagan

mean(residuals(modelo_interaccion2))  # Media de los residuos


# Modelo 3: modelo_interaccion3
modelo_interaccion3 <- lm(Cortisol_sang ~ Smoking + AROPE_cat, data = BDUNI_NA_SANG)

# Verificación de los supuestos para modelo_interaccion3
#hist(residuals(modelo_interaccion3))  # Histograma de los residuos
qqnorm(residuals(modelo_interaccion3))  # Q-Q plot de los residuos
qqline(residuals(modelo_interaccion3))  # Línea de referencia para normalidad
shapiro.test(residuals(modelo_interaccion3))  # Prueba de Shapiro-Wilk

acf(residuals(modelo_interaccion3))  # ACF: autocorrelation function
pacf(residuals(modelo_interaccion3)) # PACF: partial autocorrelation function
durbinWatsonTest(modelo_interaccion3)  # Prueba de autocorrelación Durbin-Watson

plot(fitted(modelo_interaccion3), residuals(modelo_interaccion3))  # Gráfico de residuos vs valores ajustados
bptest(modelo_interaccion3)  # Prueba de homocedasticidad Breusch-Pagan

mean(residuals(modelo_interaccion3))  # Media de los residuos
# Modelo 4: modelo_interaccion4
modelo_interaccion4 <- lm(Cortisol_sang ~ AROPE_cat + Smoking + IUD , data = BDUNI_NA_LOG_SANG)

# Verificación de los supuestos para modelo_interaccion4
#hist(residuals(modelo_interaccion4))  # Histograma de los residuos
qqnorm(residuals(modelo_interaccion4))  # Q-Q plot de los residuos
qqline(residuals(modelo_interaccion4))  # Línea de referencia para normalidad
shapiro.test(residuals(modelo_interaccion4))  # Prueba de Shapiro-Wilk

ºacf(residuals(modelo_interaccion4))  # ACF: autocorrelation function
pacf(residuals(modelo_interaccion4)) # PACF: partial autocorrelation function

durbinWatsonTest(modelo_interaccion4)  # Prueba de autocorrelación Durbin-Watson

plot(fitted(modelo_interaccion4), residuals(modelo_interaccion4))  # Gráfico de residuos vs valores ajustados
bptest(modelo_interaccion4)  # Prueba de homocedasticidad Breusch-Pagan
```

```
mean(residuals(modelo_interaccion4))  # Media de los residuos
summary(modelo_interaccion4)

#PREDICCIONES
# Predicciones en nuevos datos (si dispones de ellos)
#predict con el modelo final, y new data
#comparar predicciones del modelo con mi base de datos anterior, para evaluar la predicción del modelo
#para hacer predicciones pued hacerlo con una nueva bbdd, por ej 4 o 5 datos con solo las variables del modelo

# Supongamos que esta es la nueva base de datos con los valores de las variables
nueva_base2 <- data.frame(
  AROPE_cat = c(0, 0,0,0,1,1,1,1),      # Ejemplo de valores para AROPE_cat
  Smoking = c(0, 0,1,1,0,0,1,1),        # Ejemplo de valores para Smoking
  IUD = c(0, 1,0,1,0,1,0,1)
)

nueva_base <- as.data.frame(lapply(nueva_base2, as.factor))
nuevas_predicciones <- predict(modelo_interaccion4, newdata = nueva_base)
print(nuevas_predicciones)

str(BDUNI_NA_LOG_SANG)
nueva_base$AROPE_cat <- factor(nueva_base$AROPE_cat)

#parte de PLS

install.packages("pls")
library(pls)

# Definir variables predictoras (X) y la respuesta (Y)
X <- as.matrix(BDUNI_log_sang[, -which(colnames(BDUNI_log_sang) == "Cortisol_sang")])
# Todas menos la variable objetivo
Y <- BDUNI_log_sang$Cortisol_sang # Variable de respuesta

# Ajustar el modelo PLS usando fórmula
modelo_pls <- plsr(Cortisol_sang ~ ., data = BDUNI_log_sang, validation = "CV")
# Validación cruzada: selecciona el número óptimo de componentes
validationplot(modelo_pls, val.type = "MSEP")

# Resumen del modelo
summary(modelo_pls)
```