

Análisis de la relación entre los biomarcadores asociados al estrés y variables sociodemográficas para analizar las diferencias entre grupos étnicos

Jone Renteria Aguirregabiria

Máster en Bioinformática y Bioestadística
Análisis de datos y técnicas de clustering

Dr. Daniel Fernández Martínez

Dr. Marc Maceira Duch

© Jone Renteria

Reservados todos los derechos. Está prohibido la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la impresión, la reprografía, el microfilme, el tratamiento informático o cualquier otro sistema, así como la distribución de ejemplares mediante alquiler y préstamo, sin la autorización escrita del autor o de los límites que autorice la Ley de Propiedad Intelectual.

FICHA DEL TRABAJO FINAL

Título del trabajo:	<i>Análisis de la relación entre los biomarcadores asociados al estrés y variables sociodemográficas para analizar las diferencias entre grupos étnicos</i>
Nombre del autor:	<i>Jone Renteria Aguirregabiria</i>
Nombre del consultor/a:	<i>Dr. Daniel Fernández Martínez</i>
Nombre del PRA:	<i>Dr. Marc Maceira Duch</i>
Fecha de entrega (mm/aaaa):	<i>01/2021</i>
Titulación::	<i>Máster en Bioinformática y Bioestadística</i>
Área del Trabajo Final:	<i>Análisis de datos y técnicas de clustering</i>
Idioma del trabajo:	<i>Español</i>
Palabras clave	<i>Oxitocina, Cortisol, modelo de regresión</i>

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

Diferentes estudios relacionan el estrés, medido mediante biomarcadores como la oxitocina y el cortisol, con diversas variables sociodemográficas. Además, estudios plantean que el trauma histórico sufrido por algunas etnias podría afectar a los vínculos sociales y generar una respuesta diferente a situaciones de estrés. En este trabajo se ha querido relacionar el estrés con diferentes variables combinando estudios previamente publicados con datos recogidos en un proyecto piloto de la Universidad de Maryland (UMD, EEUU). En una primera parte, se ha realizado un análisis exploratorio completo utilizando las observaciones de la literatura (sin incluir la etnia), y posteriormente, se ha definido el modelo que mejor ajuste ha mostrado para cada biomarcador. Se ha visto que por ejemplo las variables que miden el ritmo cardíaco son significativas, lo que puede ayudar a mejorar los protocolos de recogida de datos actuales que omiten algunas de estas variables. Otro de los objetivos es analizar la hipótesis de la etnia y ver si la variable es significativa para medir los cambios en los niveles de estrés medidos mediante los biomarcadores mencionados como respuesta a distintos estímulos. Para ello, se ha querido aplicar el modelo sobre el conjunto de datos de la UMD, pero este último análisis se ha visto perjudicado por la pandemia del SARS-CoV-2/COVID-19 actual. Sin embargo, el trabajo muestra el procedimiento de una manera teórica para poder aplicarlo cuando el conjunto de datos disponga de más observaciones.

Abstract (in English, 250 words or less):

Biomarkers such as oxytocin and cortisol are biological measures to quantify the stress level of an individual, which is related with several sociodemographic factors in many scientific publications. Some studies hypothesize that the response to stressful situations can vary depending on the ethnicity of each person, being the ethnicities that have suffered from historic trauma the most vulnerable ones to overcome those stressful situations and to create social bonds. In order to analyze the relation between the stress with different variables, this work combines previously published articles with data from a pilot study from the University of Maryland (UMD). An exploratory data analysis for each of the biomarkers using published databases, without the ethnicity variable is followed by a regression analysis to find the model that best fits the data. The outcomes show that variables like the heart rate of the individual are significant covariates, so adding those to the existing data collection protocol would improve the procedure and make it more suitable. Another objective of the present work is to analyze the significance of the ethnicity variable when measuring the changes in the stress, based on the biomarker's level. This goal has been affected by the current SARS-CoV-2/COVID-19 pandemic, and it has hinder the possibility to apply the regression models to the real project's data. Nevertheless, this work shows the process in a theoretical way. The application of the regression models to check the significance of the ethnicity to explain changes in stress is expected to be completed as soon as more observations are included in the database.

Índice

1. Introducción	1
1.1 Contexto y justificación del Trabajo	1
1.2 Objetivos del Trabajo.....	3
1.2.1 Objetivos generales	3
1.2.2 Objetivos específicos.....	3
1.3 Enfoque y método seguido.....	4
1.3.1 Preparación de los datos	4
1.3.2 Descriptiva de los datos.....	4
1.3.3 Definición de los modelos	4
1.3.4 Predicción y análisis de los residuos del modelo.....	4
1.3.5 Interacción entre las variables	5
1.3.6 Reducción/simplificación del modelo	5
1.3.7 Analizar la significación de las variables en los modelos finales	5
1.3.8 Incorporación de la variable etnia.....	5
1.4 Planificación del Trabajo.....	6
1.4.1 Recursos necesarios	6
1.4.2 Tareas	7
1.4.3 Calendario.....	7
1.4.4 Hitos.....	9
1.4.5 Análisis de riesgos.....	9
1.5 Breve sumario de productos obtenidos	10
1.6 Breve descripción de los otros capítulos de la memoria.....	11
2. Metodología.....	12
2.1. Planteamiento inicial y generación de la base de datos.....	12
2.2. Descriptiva de los datos.....	12
2.3. Biomarcador I: Oxitocina	15
2.3.1 Variable respuesta	15
2.3.2 Valores faltantes en el conjunto de datos.....	17
2.3.3 Variables predictoras.....	19
2.3.4 Análisis de la correlación de variables.....	24
2.3.5 Modelo	26
2.3.5.1 Normalidad de los residuos:.....	29
2.3.5.2 Homocedasticidad/heterocedasticidad:	29
2.3.5.3 Autocorrelación:.....	30
2.3.5.4 Multicolinealidad:.....	30
2.3.6 Conclusión modelo Oxitocina	30
2.4 Biomarcador II: Cortisol	31
2.4.1 Variable respuesta	31
2.4.2 Valores faltantes en el conjunto de datos.....	33
2.4.3 Variables predictoras.....	35
2.4.4 Análisis de la correlación de variables.....	41
2.4.5 Modelo	45
2.4.5.1 Propuesta 1.....	45
2.4.5.2 Propuesta 2.....	48
2.4.5.2.1 Sangre	48
Modelo sangre - cortisol.....	52
Conclusión modelo y comparación.....	55
2.4.5.2.2 Saliva	56

Modelo saliva - cortisol.....	60
Conclusión modelo y comparación.....	63
2.4.6 Conclusión modelo cortisol	63
2.5 Aplicación de los modelos	64
2.6 Repositorio online.....	65
3. Conclusiones.....	66
4. Bibliografía	67
Anexos.....	IV

Lista de figuras

Figura 1: cronograma del proyecto.....	8
Figura 2: boxplot de la variable oxitocina tras aplicar un estímulo sobre el participante, donde se muestran los valores de la media, mediana Q1, Q3, min y max	15
Figura 3: distribución de la variable respuesta que mide el nivel de oxitocina tras aplicar un estímulo sobre el participante (oxt.post). Arriba a la izquierda, histograma de la distribución original. Arriba a la derecha, gráfico QQ de los datos originales. Los gráficos de abajo muestran histogramas de la distribución de la variable en caso de aplicar la transformación logarítmica o de raíz cuadrada a los datos.....	17
Figura 4: valores faltantes en el conjunto de datos data.oxt obtenido mediante la función agrgr del paquete VIM. Proporción de valores faltantes en variable numéricas (orden ascendente en la dirección marcada) para tres combinaciones	18
Figura 5: valores faltantes del conjunto de datos data.oxt eliminando la variable med.dos, donde hay 46 observaciones completas, 32 donde hay valores faltantes en tres variables y 6 observaciones con valores faltantes en dos variables.....	18
Figura 6: boxplots con los valores de la media, mediana, Q1, Q3, min y max para las variables del ritmo cardiaco (hr.bas-izq.-, hr.post –centro-) y oxitocina previa al estímulo (oxt.pre-dch.-) separados según si no se aplica ningún estímulo o se aplica un estímulo físico sobre el participante	20
Figura 7: boxplot de la variable edad para ambos grupos de estímulos del conjunto de datos data.oxt. Se muestran los datos numéricos del mínimo, máximo, media, mediana, Q1 y Q3 ...	22
Figura 8: análisis de la normalidad para la variable edad (arriba izquierda), oxitocina pre-estímulo (oxt.pre, arriba a la derecha), ritmo cardiaco previo al estímulo (hr.bas, abajo a la izquierda) y ritmo cardiaco post estímulo (hr.post, abajo a la derecha). Para cada una se muestra la distribución original mediante histograma y gráfico QQ, e histograma con transformación log y sqrt	23
Figura 9: relación lineal entre la variable predictora de oxitocina previa al estímulo (oxt.pre) y la variable respuesta de oxitocina post-estímulo (oxt.post)	24
Figura 10: coeficientes de correlación del conjunto de datos data.oxt visualizados de forma gráfica. A la izquierda, todos los coeficientes, y a la derecha, visualización de los coeficientes significativos al 5%. Círculo más grande y oscuro, mayor correlación.....	25
Figura 11: heatmap para el análisis de la correlación entre las variables del conjunto de datos data.oxt. Los rectángulos rojos identifican los coeficientes de correlación más cercanos a uno (más intensidad de rojo mayor correlación), y los rectángulos azules, menor correlación (mayor intensidad de azul menor correlación)	26
Figura 12: distribución de los residuos del modelo mod.oxt2 (transformación doble log en las variables numéricas) para predecir el nivel de oxitocina tras aplicar un estímulo. Gráfico de linealidad (arriba izq.), homocedasticidad (arriba dcha.), normalidad (abajo izq.) y puntos outliers o influyentes (abajo dcha.).....	28
Figura 13: boxplot de la variable cortisol tras aplicar un estímulo sobre el participante, donde se muestran los valores de la media, mediana Q1, Q3, min y max, utilizando el conjunto de datos completo	32
Figura 14: distribución de la variable respuesta que mide el nivel de cortisol tras aplicar un estímulo sobre el participante (co.post). Arriba a la izquierda, histograma de la distribución original. Arriba a la derecha, gráfico QQ de los datos originales. Los gráficos de abajo muestran histogramas de la distribución de la variable en caso de aplicar la transformación logarítmica o de raíz cuadrada a los datos. Análisis del conjunto de datos completo	33
Figura 15: valores faltantes conjunto de datos cortisol para las variables numéricas, obtenido con la función agrgr del paquete VIM. Proporción de valores faltantes (orden ascendente en la dirección marcada) para tres combinaciones	34

- Figura 16:** valores faltantes del conjunto de datos del cortisol data.co, donde hay 51 observaciones completas, 32 donde hay valores faltantes en dos variables y 1 observaciones con valores faltantes en tres variables. Hay 67 valores faltantes en total 34
- Figura 17:** gráfico de barras de las variables categóricas co.res, gender y stimulus.type, que miden si el participante muestra un cambio o no en el nivel de cortisol tras el estímulo aplicado, el género del participante y el tipo de estímulo aplicado. Se utiliza el conjunto de datos del cortisol completo. 36
- Figura 18:** boxplots con los valores de la media, mediana, Q1, Q3, min y max para las variables numéricas del conjunto de datos del cortisol. Fila de arriba, izquierda a derecha: nivel de cortisol previo al estímulo, índice de reacción al cortisol y dosis del medicamento. Fila de abajo, izq. a dcha: ritmo cardiaco antes y después del estímulo aplicado en cada caso. Todos los gráficos están separados el tipo de estímulo que se aplique sobre el participante: ninguno, psicológico o físico 37
- Figura 19:** boxplot de la variable edad para cada tipo de estímulo del conjunto de datos data.co completo. Se muestran los datos numéricos del mínimo, máximo, media, mediana, Q1 y Q3 sobre el gráfico 39
- Figura 20:** análisis de la normalidad. Fila arriba, izq. a dcha: variable edad, dosis ingerida, reacción del cortisol. Fila abajo, izq. a dcha: nivel de cortisol pre-estímulo, ritmo cardiaco post estímulo y ritmo cardiaco previo al estímulo Para cada variable se muestra la distribución original mediante histograma y gráfico QQ, e histograma con transformación log y sqrt 40
- Figura 21:** relación lineal entre la variable respuesta que mide el nivel de cortisol post estímulo (co.post) y la variable predictora del cortisol previo al estímulo (co.pre), donde se observa una gran correlación entre ambas variables 44
- Figura 22:** mapa de calor, heatmap para visualizar la correlación entre las variables del conjunto de datos data.co utilizando el conjunto de datos completo. Los rectángulos rojos identifican los coeficientes de correlación más cercanos a uno (más intensidad de rojo mayor correlación), y los rectángulos azules, menor correlación (mayor intensidad de azul menor correlación) 44
- Figura 23:** distribución de los residuos del modelo mod.co.p1 (transformación doble log sobre todas las covariables y la variable dependiente) para predecir el nivel de cortisol tras aplicar un estímulo. Gráfico de linealidad (arriba izq.), homocedasticidad (arriba dcha.), normalidad (abajo izq.) y puntos outliers o influyentes (abajo dcha.) 47
- Figura 24:** boxplot de la variable cortisol tras aplicar un estímulo sobre el participante, donde se muestran los valores de la media, mediana Q1, Q3, min y max, utilizando el conjunto de datos con las mediciones en sangre 48
- Figura 25:** distribución de la variable respuesta que mide el nivel de cortisol tras aplicar un estímulo sobre el participante (co.post). Arriba a la izquierda, histograma de la distribución original. Arriba a la derecha, gráfico QQ de los datos originales. Los gráficos de abajo muestran histogramas de la distribución de la variable en caso de aplicar la transformación logarítmica o de raíz cuadrada a los datos. Conjunto de datos con las mediciones en sangre 49
- Figura 26:** boxplots con los valores de la media, mediana, Q1, Q3, min y max para las variables numéricas del conjunto de datos con las mediciones de la sangre. Fila de arriba: variable edad y nivel de cortisol previo a la aplicación del estímulo. Fila abajo: nivel de reacción frente a los estímulos y dosis ingerida de los participantes. Todas las observaciones pertenecen al tipo de estímulo psicológico 50
- Figura 27:** análisis de la normalidad. Fila arriba, izq. a dcha: variable edad y dosis ingerida. Fila abajo, izq. a dcha: nivel de cortisol previo al estímulo y reacción del cortisol frente a los estímulos. Conjunto de datos del cortisol medido en sangre. Para cada variable se muestra la distribución original mediante histograma y gráfico QQ, e histograma con transformación log y sqrt 51
- Figura 28:** mapa de calor (heatmap) a partir de los coeficientes de correlaciones para las variables del conjunto de datos del cortisol medido en sangre. Los rectángulos rojos identifican

los coeficientes de correlación más cercanos a uno (más intensidad de rojo mayor correlación), y los rectángulos azules, menor correlación (mayor intensidad de azul menor correlación)	52
Figura 29: distribución de los residuos del modelo mod.co.sngr3 (transformación logarítmica de la variable respuesta). Gráfico de linealidad (arriba izq.), homocedasticidad (arriba dcha.), normalidad (abajo izq.) y puntos outliers o influyentes (abajo dcha.)	53
Figura 30: boxplot de la variable cortisol tras aplicar un estímulo sobre el participante, donde se muestran los valores de la media, mediana Q1, Q3, min y max, utilizando el conjunto de datos con las mediciones en la saliva.....	57
Figura 31: distribución de la variable respuesta que mide el nivel de cortisol tras aplicar un estímulo sobre el participante (co.post). Arriba a la izquierda, histograma de la distribución original. Arriba a la derecha, gráfico QQ de los datos originales. Los gráficos de abajo muestran histogramas de la distribución de la variable en caso de aplicar la transformación logarítmica o de raíz cuadrada a los datos. Conjunto de datos del cortisol con mediciones de la saliva	58
Figura 32: boxplots con los valores de la media, mediana, Q1, Q3, min y max para las variables numéricas del conjunto de datos con las mediciones de la saliva según el estímulo aplicado. Fila de arriba: variable edad, nivel de cortisol previo y reacción del cortisol. Fila de abajo: niveles del ritmo cardiaco (hr.bas y hr.post).....	58
Figura 33: análisis de la normalidad. De izq. a dcha: variables age, co.pre y co.reac. Conjunto de datos del cortisol medido en la saliva. Para cada variable se muestra la distribución original mediante histograma y gráfico QQ, e histograma con transformación log y sqrt.....	59
Figura 34: mapa de calor (heatmap) a partir de las correlaciones para las variables del conjunto de datos del cortisol medido en la saliva. Los rectángulos rojos identifican los coeficientes de correlación más cercanos a uno (más intensidad de rojo mayor correlación), y los rectángulos azules, menor correlación (mayor intensidad de azul menor correlación).	60
Figura 35: distribución de los residuos del modelo mod.co.slv2 (transformación logarítmica de la variable respuesta y las covariables). Gráfico de linealidad (arriba izq.), homocedasticidad (arriba dcha.), normalidad (abajo izq.) y puntos outliers o influyentes (abajo dcha.).....	62

Lista de tablas

Tabla 1: listado de tareas numeradas, con el tiempo estimado y la fecha límite para realizarla. Cada tarea se asocia al objetivo general y específico definido en el segundo apartado de este entregable.....	7
Tabla 2: hitos para la elaboración del proyecto.....	9
Tabla 3: descriptiva datos base de datos	13
Tabla 4: niveles de las variables categóricas.....	14
Tabla 5: descriptiva numérica de la variable respuesta oxt.post (nivel de oxitocina tras aplicar un estímulo sobre el participante)	16
Tabla 6: número de observaciones completas e incompletas del conjunto de datos data.oxt, y descripción de qué variables tienen valores faltantes.....	19
Tabla 7: descriptiva numérica de las variables oxt.pre, hr.bas y hr.post, tanto de forma general como separandolas por el tipo de estímulo aplicado sobre ellas. Se recogen valores generales (min, max, media, mediana, Q1, Q3) y valores de las medidas de dispersión de cada una (varianza, rango, IQR).....	21
Tabla 8: descriptiva numérica variable edad, donde se recoge el valor mínimo, máximo, cuantiles, media, y valores de las medidas de dispersión (varianza, rango, IQR).....	22
Tabla 9: correlación de las variables del conjunto de datos data.oxt aplicando el método de Spearman	24
Tabla 10: resultados del modelo de regresión mod.oxt2 para predecir el nivel de oxitocina post aplicación de un estímulo sobre un participante, con cuatro covariables: age, oxt.pre y hr.bas trasnsfromadas logarítmicamente y el tipo de estímulo	27

Tabla 11: descriptiva numérica de la variable respuesta co.post (nivel de cortisol tras aplicar un estímulo sobre el participante)	32
Tabla 12: tabla de frecuencias de las variables categóricas del conjunto de datos del biomarcador cortisol donde las muestras se han recogido en la saliva. Entre paréntesis el %. *Existe un valor NA para el tipo de estímulo “ninguno”	36
Tabla 13: tabla de frecuencias de las variables categóricas del conjunto de datos del biomarcador cortisol donde las muestras se han recogido en la sangre. Entre paréntesis el %.	36
Tabla 14: descriptiva numérica de las variables co.pre, co.reac, med.dos, hr.bas y hr.post, tanto de forma general como separandolas por el tipo de estímulo aplicado sobre ellas. Se recogen valores generales (min, max, media, mediana, Q1, Q3) y valores de las medidas de dispersión de cada una (varianza, rango, IQR)	38
Tabla 15: descriptiva numérica variable edad separada por el tipo de estímulo aplicado y de forma general, donde se recoge el valor mínimo, máximo, cuantiles, media, y valores de las medidas de dispersión (varianza, rango, IQR).....	39
Tabla 16: matriz de correlación entre las variables que componen el conjunto de datos del cortisol (data.co) general, aplicando el método de Spearman	42
Tabla 17: resultados del modelo de regresión para predecir el nivel de cortisol tras aplicar el estímulo en el participante, con el logaritmo de las covariables co.pre y co.reac como predictores del nivel de cortisol	46
Tabla 18: descriptiva numérica de la variable respuesta co.post (nivel de cortisol tras aplicar un estímulo sobre el participante) para el conjunto de datos con mediciones en sangre	49
Tabla 19: descriptiva numérica de las covariables co.pre, co.reac, med.dos y age de forma general (estímulo psicológico). Se recogen valores generales (min, max, media, mediana, Q1, Q3) y valores de las medidas de dispersión de cada una (varianza, rango, IQR). Conjunto de datos del cortisol con mediciones en sangre	50
Tabla 20: matriz de correlación para las variables del conjunto de datos del cortisol medido en sangre.....	51
Tabla 21: resultados del modelo de regresión, variables co.pre, age, co.reac y med.dos como predictores del nivel de cortisol post situación de estrés.....	53
Tabla 22: descriptiva numérica de la variable respuesta co.post (nivel de cortisol tras aplicar un estímulo sobre el participante) para el conjunto de datos con mediciones en la saliva	57
Tabla 23: descriptiva numérica de las covariables co.pre, co.reac, age y ritmos cardiacos de forma general. Se recogen valores generales (min, max, media, mediana, Q1, Q3) y valores de las medidas de dispersión de cada una (varianza, rango, IQR). Conjunto de datos del cortisol con mediciones de saliva	59
Tabla 24: matriz de correlaciones para las variables del conjunto de datos del cortisol medido en la saliva	60
Tabla 25: resultados del modelo de regresión, logarítmico de las variables co.pre y co.reac como predictores del nivel de cortisol tras la aplicación del estímulo	61

Lista de ecuaciones

Ecuación 1: planteamiento inicial del modelo para predecir el nivel de oxitocina tras aplicar un estímulo sobre un participante con las covariables numéricas y la variable dependiente transformadas logarítmicamente	27
Ecuación 2: ecuación final incluyendo los coeficientes de cada covariable para describir el modelo mod.oxt2 y predecir el nivel de oxitocina tras aplicar un estímulo sobre el participante, transformando logarítmicamente las covariables numéricas y la variable respuesta	31
Ecuación 3: planteamiento inicial modelo con mejores resultados en la propuesta 1 para predecir el nivel de cortisol tras la aplicación de un estímulo sobre el participante. Conjunto de datos completo del cortisol, data.co. Variable dependiente y covariables transformadas logarítmicamente	46
Ecuación 4: planteamiento inicial del modelo mod.co.sngr3 utilizando el conjunto de datos del cortisol para las mediciones en sangre y predecir el nivel de cortisol tras la aplicación del estímulo. Variable dependiente co.post, transformada logarítmicamente.	52
Ecuación 5: ecuación final incluyendo los coeficientes de cada covariable para describir el modelo mod.co.sngr3 y predecir el nivel de cortisol tras aplicar un estímulo sobre el participante, utilizando el conjunto de datos del cortisol para las mediciones obtenidas en la sangre. Transformación logarítmica de la variable respuesta co.post.	55
Ecuación 6: planteamiento inicial del modelo mod.co.slv2 utilizando el conjunto de datos del cortisol para las mediciones en saliva y predecir el nivel de cortisol tras la aplicación de un estímulo sobre el participante. Transformación logarítmica de la variable respuesta y las covariables.	61
Ecuación 7: ecuación final incluyendo los coeficientes de cada covariable para describir el modelo mod.co.slv2 y predecir el nivel de cortisol tras aplicar un estímulo sobre el participante, utilizando el conjunto de datos del cortisol para las mediciones obtenidas de la saliva. Transformación de la variable respuesta co.post y las covariables.	63

1. Introducción

1.1 Contexto y justificación del Trabajo

El estrés en una persona está influenciado, entre otras cosas, por diversos factores sociodemográficos (Irizar y Haro 2017), y como consecuencia, éstos tienen un impacto negativo en las competencias sociales y académicas actuales tanto a nivel personal como familiar (Cabrera et al. 2016). Existe la hipótesis de que el estrés, comúnmente sufrido por gente de bajos recursos, afecta de forma diferente según el grupo étnico (Anderson et al. 2004; Hwang y Ting 2008; Gallo et al. 2009; Panchang et al. 2016; Boileau et al. 2019), y esto podría ser a causa del estrés acumulado en ciertas etnias debido a la represión racial sufrida durante las últimas décadas (Goosby y Heidbrink 2013). Además, es posible que el trauma histórico haya tenido un impacto intergeneracional negativo en otros aspectos del desarrollo humano, como por ejemplo en la capacidad de generar vínculos sociales (Cabrera et al. 2016; Halloran 2019). Un ejemplo de ello es la comparación entre la población afroamericana y la latina en Estados Unidos, donde el entorno social en el que viven los latinos tiene una influencia positiva en los resultados académicos, pese a vivir en hogares más vulnerables en relación a su economía y educación, a diferencia de los afroamericanos, donde el entorno social hace que sus competencias académicas sean peores (Cabrera et al. 2016).

A consecuencia del estrés, también se podrán generar problemas psicológicos y físicos y estos últimos podrán influenciar en el sistema inmune, cardiovascular, endocrino (derivando en enfermedades como la diabetes), gastrointestinal o el sistema nervioso central de las personas que lo padecen (Anderson 1998; Öhman et al. 2007; Salleh 2008; Kronenberg et al. 2017). Para medir el estrés en una persona de forma cuantitativa y a través de muestras biológicas (serológicas o de saliva), se analizan los valores obtenidos del cortisol. Este biomarcador, es un glucocorticoide que se produce en las glándulas suprarrenales (Juster, McEwen, y Lupien 2010) y que se libera como respuesta al estrés. Cuando un estímulo estresante se repite de forma crónica, el cortisol se mantiene en niveles más elevados durante un periodo de tiempo prolongado, y además tiene la capacidad de mantener los niveles elevados aun cuando el estímulo que ha generado la situación de estrés ha desaparecido (Lee, Kim, y Choi 2015).

Tal y como se ha comentado previamente, la capacidad de generar vínculos sociales puede estar relacionado con el nivel de estrés que una persona sufre en su día a día. Es decir, cuanto mayor es el nivel de estrés con el que convive esa persona, la capacidad de generar lazos sociales entre individuos es menor. Se ha demostrado que la oxitocina (neuropeptido que se sintetiza en el núcleo supraóptico y en el núcleo paraventricular del hipotálamo), promueve las interacciones sociales positivas y tiene un efecto ansiolítico y anti-estresante, atenuando por ello los niveles de estrés psicológicos y de conducta en una persona (Sue Carter 1998; Kumsta y Heinrichs 2013).

En este estudio se quiere analizar a través de los valores biológicos del cortisol y la oxitocina el modo en el que estos biomarcadores varían como respuesta a un estímulo estresante según la etnia, combinándolos además con otros factores socioeconómicos y sociodemográficos. Como se han observado que existen diferencias entre etnias en relación a su entorno, se quiere analizar si la variable etnia tiene un efecto significativo sobre los biomarcadores. De este modo, podríamos analizar las consecuencias que tiene hoy en día la represión racial sufrida en determinadas etnias en su vida cotidiana, y ver cómo ha afectado esto a la hora de generar vínculos sociales.

Tras llevar a cabo un análisis del estado del arte, se ha observado que la relación entre el estrés (medido mediante el cortisol) con las posteriores consecuencias en la salud de cada individuo se han analizado de forma extendida (Coleman et al. 2016; National Center for Health Statistics 2017). Existen también estudios que vinculan los cambios en los niveles de cortisol con la etnia (Boileau et al. 2019), y los que además los relacionan con factores sociales, como por ejemplo la educación (Bennett, Merritt, y Wolin 2004). Algunos artículos asocian el estrés regulado mediante los niveles de cortisol con los niveles de oxitocina (Alley et al. 2019) y otros con la capacidad de esta hormona para generar vínculos sociales (Heinrichs et al. 2003; Cardoso et al. 2013). Otros, asocian los niveles de oxitocina elevados con una capacidad de respuesta mejorada frente a una situación de estrés (Kubzansky et al. 2012). Finalmente, a nivel familiar, también hay algunos estudios publicados que estiman los niveles de cortisol en bebés recién nacidos en familias de bajos recursos y de origen mexicano en Estados Unidos (Luecken et al. 2015), o que miden los niveles del cortisol para ayudar a mitigar las consecuencias psicológicas y de salud generadas a consecuencia del estrés en madres de bajos recursos en zonas rurales alemanas (Bischoff et al. 2019). En estos dos últimos casos, únicamente se tiene en cuenta la figura materna y el bebé.

Sin embargo, tal y como se ha comentado, en el presente trabajo además de medir la significatividad de la etnia sobre el cortisol, también se analizará el efecto de la variable etnia sobre la hormona oxitocina, ya que ambas (cortisol y oxitocina) pueden influenciar en la vida cotidiana de las personas en relación al estrés y a la capacidad de generar vínculos sociales, respectivamente. Además, se utilizarán datos de familias completas (madre, padre y bebé), por lo que se podrá observar el efecto de la etnia sobre un núcleo familiar completo. Según los expertos con los que se ha trabajado, el hecho de haber obtenido datos de los padres, especialmente para las familias de bajos recursos (como es el caso de nuestra muestra) hace que el estudio sea más relevante, ya que son pocos los estudios que incluyen la figura paterna, limitando el análisis de éste en el desarrollo infantil o el ambiente en el hogar.

Para llevar a cabo este análisis, se generará un modelo estadístico (en principio lineal) para cada biomarcador con el objetivo de observar el efecto que las diferentes variables dependientes (tanto factores sociodemográficos, socioeconómicos, biológicos, etc.) ejercen sobre ellas. Este estudio actual se considera piloto, dado que los resultados de este estudio podrían incluirse en una propuesta de proyecto de colaboración entre la Universidad de Maryland y el *National Institute of Health* (NIH), lo que ayudaría a seguir con la presente investigación con más participantes en un futuro cercano. Además, los resultados preliminares obtenidos en el presente análisis ayudarán a modificar el protocolo actual de visita a los hogares para la obtención de muestras y su posterior análisis.

En las siguientes subsecciones del primer apartado de la memoria se describen los objetivos del trabajo (Sección 1.2), el enfoque y el método seguido (Sección 1.3), la planificación para la correcta elaboración del proyecto (Sección 1.4) y un breve resumen de los productos que se quieren obtener al finalizarlo (Sección 1.5). Finalmente, en la subsección 1.6, se resumen los próximos capítulos de la memoria.

1.2 Objetivos del Trabajo

Los objetivos de este proyecto se plantean a continuación:

1.2.1 Objetivos generales

1. Generar un modelo por cada biomarcador, en el cual la variable respuesta sea el valor del biomarcador y los factores sociodemográficos sean las variables explicativas. Analizar el efecto de la variable etnia.
2. Mejorar el protocolo actual de visita a los hogares para la recopilación de datos, optimizando el cuestionario actual de los participantes e incluyendo únicamente aquellas variables estadísticamente significativas.

1.2.2 Objetivos específicos

1. Generar un modelo por cada marcador biológico, en el cual el biomarcador sea la variable respuesta y relacionarla mediante inferencia estadística con otros factores socioeconómicos, demográficos y diferentes estímulos (expresadas como variables explicativas) para ver el efecto que éstos tienen sobre los biomarcadores y por lo tanto, con el estrés. Se llevará a cabo sobre los datos obtenidos en la literatura.
2. Aplicar los modelos generados a partir de los datos de la literatura sobre el conjunto de datos perteneciente al estudio piloto comenzado en 2018 en la Universidad de Maryland, con una muestra más pequeña. Testear y aplicar los modelos, añadiendo la variable explicativa etnia.
3. Analizar si la etnia es una variable significativa (añadiéndola como variable explicativa en cada modelo) para los valores esperados de los marcadores biológicos y observar su efecto sobre cada uno de ellos.
4. Ver si añadir la variable etnia mejora el modelo (bondad de ajuste, R^2) para la muestra actual.
5. Generar un script en R al que únicamente haya que introducirle una base de datos para que observe el efecto de la etnia sobre los biomarcadores.

1.3 Enfoque y método seguido

En los siguientes puntos se detalla la metodología que se plantea para llevar a cabo el trabajo. Una vez generada la base de datos definitiva a partir de los datos obtenidos de artículos previamente publicados (Tas et al. 2018; Ooishi et al. 2017), ésta se cargará en el software estadístico R. Los pasos que se esperan llevar a cabo son los siguientes:

1.3.1 Preparación de los datos

- Datos faltantes. Asegurar la misma nomenclatura en todos ellos. Se valorará la posibilidad de llevar a cabo una imputación de los datos faltantes, pero en las variables donde haya un porcentaje alto de *missings*, se trabajará con observaciones completas, filtrando aquellos valores *NA*, y sin imputarlos.
- Posibilidad de transformar alguna variable numérica en categórica en caso de que a la hora de plantear el modelo se considere necesario. Además, se analizarán y valorarán las posibles transformaciones de las variables en el caso de que alguna de las condiciones necesarias para el modelo no se cumpla. Por ejemplo, transformar las variables a una escala logarítmica o realizar la transformación *Box-cox* sobre la variable respuesta.

1.3.2 Descriptiva de los datos

- Resumen general de los datos. Estadística descriptiva (media, desviación estándar y número de participantes). Acompañar estos análisis con figuras (*scatterplots*, *boxplots* e *histogramas* por ejemplo).
- En los gráficos: observar si hay valores *outliers* (también comprobarlo numéricamente), y considerar si se deberían eliminar del estudio / ver la distribución de los datos...etc.
- Analizar las correlaciones entre las variables para evitar multicolinealidad.

1.3.3 Definición de los modelos

- Generar un modelo (en principio lineal) donde la variable respuesta sea la predicción de la hormona oxitocina (tras aplicar un estímulo sobre un individuo) según las diferentes variables del conjunto de datos.
- Llevar a cabo el mismo proceso donde la variable respuesta sea el cortisol. Se intentará que las variables explicativas sean simétricas a las variables del modelo de la oxitocina.

1.3.4 Predicción y análisis de los residuos del modelo

- Estimar β (en principio, por mínimos cuadrados ordinarios). Puntual o por intervalos de confianza. El modelo predictivo que se planteará en un principio será lineal.
- Analizar y visualizar los residuos de los modelos y su ajuste: normalidad (Shapiro-Wilk y gráficamente), homocedasticidad (tests y gráficamente), autocorrelación (test Durbin-Watson) y linealidad de los residuos.
- En caso de llevar a cabo transformaciones en las variables (tal y como se menciona en el apartado 1.3.1), volver a ejecutar los modelos y analizar los criterios mencionados previamente (normalidad, homocedasticidad, autocorrelación y linealidad) de los residuos.

1.3.5 Interacción entre las variables

- Observar si el modelo I (variable respuesta oxitocina) y el modelo II (variable respuesta cortisol) se pueden mejorar añadiendo la interacción entre dos variables explicativas. Esto es, en el caso de sospechar que dos o más variables de efecto fijo o aleatorio pueden estar relacionadas, es posible añadir la interacción en los modelos. En este caso, como en el modelo I y en el modelo II las variables explicativas se intentará que sean las mismas, se añadiría en ambos modelos en caso de considerarlo necesario.
- En caso de añadir la interacción entre alguna variable, volver a analizar si los criterios en los residuos del nuevo modelo se cumplen (volver al paso 3.4).

1.3.6 Reducción/simplificación del modelo

- La reducción de los modelos se puede llevar a cabo con cualquiera de los siguientes métodos:
 - AIC, BIC o *stepwise (backward, forward o ambos)* para simplificar los modelos.
 - Contraste de modelos. Mediante test ANOVA se observarán las diferencias entre dos modelos (en caso de que se asuma la normalidad de los residuos de los mismos), donde por ejemplo, en uno de ellos una de las variables estará ausente o se asuma como hipótesis nula que la β entre dos variables explicativas sea igual. También se podrán utilizar las funciones AIC y BIC entre los modelos que cumplan las condiciones y observar con qué modelo se obtiene un valor menor. En caso de que los residuos no sean normales, se utilizará un test de permutaciones.

En caso de simplificar alguno de los modelos, volver a analizar si los criterios en los residuos del nuevo modelo se cumplen (*volver al paso 3.4*).

- Finalmente, analizar la multicolinealidad en los modelos.

1.3.7 Analizar la significación de las variables en los modelos finales

- Determinar la significación de la regresión de los modelos I y II. Se observará mediante el valor de la bondad de ajuste (R^2 ajustado) si el modelo ha mejorado y mediante el p-valor de cada variable si estas son significativas para la predicción de las hormonas (este paso puede realizarse al mismo tiempo que el paso 3.5).

1.3.8 Incorporación de la variable etnia

- Con los modelos I y II definitivos (con un valor de la bondad de ajuste que aceptemos y las variables significativas) obtenidos a partir de los datos de los datos de la literatura, repetir los pasos definidos anteriormente para la variable etnia utilizando los datos reales del estudio piloto: significatividad de la variable, no alteración de los residuos, análisis de la variación en la bondad de ajuste.

1.4 Planificación del Trabajo

El presente apartado describe la planificación del proyecto, dividida en diferentes subapartados que se detallan a continuación: primero, se describen los recursos que se esperan sean necesarios para la elaboración completa del proyecto (Sección 1.4.1), y posteriormente se elabora un listado de las tareas en relación a cada uno de los objetivos generales y específicos definidos en el apartado 1.2 junto con el tiempo que se espera sea necesario para su realización y la fecha límite de cada una de las tareas (Tabla 1 de la sección 1.4.2). En el tercer subapartado (1.4.3), se muestra un cronograma (Figura 1) que plasma de forma gráfica cada una de las tareas indicando el tiempo máximo previamente definido en la Tabla 1, y además añade puntos de referencia entre las tareas. Estos puntos, se denominan hitos y se utilizan para supervisar el progreso del proyecto, pudiendo englobar más de una tarea en cada uno de ellos. Se resumen en la Tabla 2 del subapartado 1.4.4. En el último subapartado (1.4.5), se muestran los posibles riesgos que podrían suponer un problema para completar el trabajo en la forma en la que se ha diseñado al comienzo del proyecto.

1.4.1 Recursos necesarios

Debido a las características del presente proyecto, donde se quiere analizar el efecto de ciertas muestras biológicas humanas para determinar cómo éstas varían en función de unas propiedades determinadas, es evidente que el recurso principal sea el humano: son necesarias, por una parte, las muestras biológicas de gente sometida a proyectos con humanos aprobados por comités de ética, y por otra parte, los recursos humanos necesarios para recoger las muestras y los datos de los participantes, para analizar las muestras recogidas, para hacer el análisis estadístico necesario y obtener los resultados pertinentes y también los recursos para la supervisión del proyecto durante todo el proceso. Otro recurso humano vinculado al presente proyecto es el relacionado con los diferentes investigadores que han publicado artículos y bases de datos utilizando datos de los biomarcadores cortisol y oxitocina, relacionándolos con un estado de estrés. De este último recurso se espera que tras ser contactados, dicha gente esté dispuesta a compartir sus datos para fines académicos y poder utilizarlos para generar la primera base de datos y llevar a cabo el análisis estadístico preliminar.

En relación a los recursos físicos necesarios, en este caso destacaría el equipamiento del laboratorio IDC Herzliya de Israel donde se han medido las muestras de saliva recogidas en las visitas a los hogares de Estados Unidos. Sin embargo, la situación de pandemia debido al SARS-CoV-2/COVID-19 que actualmente estamos viviendo a nivel mundial ha hecho que los equipos únicamente puedan ser utilizados para fines relacionados con el virus (al menos en este laboratorio), paralizando la medición de muestras de otros proyectos.

El recurso material para la elaboración del proyecto se compone básicamente de softwares (principalmente R¹ y también programas secundarios como Ganttproject² para algunas de las partes puntuales del proyecto), y buscadores bibliográficos médicos para generar el conjunto de datos utilizando artículos y datos previamente publicados (tales como Scopus³ o Pudmed⁴). El recurso material necesario para llevar a cabo el trabajo será básicamente una computadora.

¹ <https://www.r-project.org/>

² <https://www.ganttproject.biz>

³ <https://www.scopus.com/>

⁴ <https://pubmed.ncbi.nlm.nih.gov/>

1.4.2 Tareas

Las tareas principales definidas para llevar a cabo el proyecto se muestran en la Tabla 1 que se observa a continuación:

Tabla 1: listado de tareas numeradas, con el tiempo estimado y la fecha límite para realizarla. Cada tarea se asocia al objetivo general y específico definido en el segundo apartado de este entregable.

# Tarea	# Obj. general	# Obj. específico	Tarea	Tiempo estimado (días)	Fecha límite
1	1	1	Búsqueda bibliográfica	61	16.10.2020
2	-	-	Preparación PEC 0	13	28.09.2020
3	-	-	Preparación PEC 1	14	13.10.2020
4	1	1	Generación de la base de datos	9	15.10.2020
5	1	1	Planteamiento inicial del modelo de regresión	5	19.10.2020
6	1	1	Análisis de datos en R_I	10	26.10.2020
7	1	1	Análisis de datos en R_II	17	12.11.2020
8	1	1	Preparación PEC 2	20	15.11.2020
9	1	2	Generación base de datos – Datos piloto	6	21.11.2020
10	1	2,3,4	Análisis de datos en R_III	16	07.12.2020
11	1	1,2,3,4	Preparación PEC 3	17	17.12.2020
12	1	5	Preparación script R	14	13.12.2020
13	2	2,3,4	Mejora protocolo actual	4	19.12.2020
14	1	1,2,3,4	Conclusiones y resultados	8	21.12.2020
15	1,2	1,2,3,4	Preparación PEC 4	18	04.01.2021
16	1,2	1,2,3,4	Preparación PEC 5a	5	08.01.2021
17	1,2	1,2,3,4,5	Preparación <i>executive summary</i>	5	16.01.2021
18	1,2	1,2,3,4	Preparación PEC 5b	7	19.01.2021

1.4.3 Calendario

En la presente subsección se muestra de forma gráfica el tiempo máximo que se debe emplear en cada una de las tareas definidas en el subapartado 1.4.2, así como los hitos establecidos a lo largo del proyecto para la correcta elaboración del trabajo. El cronograma (Figura 1) se ha llevado a cabo utilizando el programa *GanttProject*. Nótese que los colores definidos en la Tabla 1 coinciden con cada tarea descrita en el gráfico. Los hitos se muestran mediante rombos verdes y los festivos utilizando columnas de color rosa.

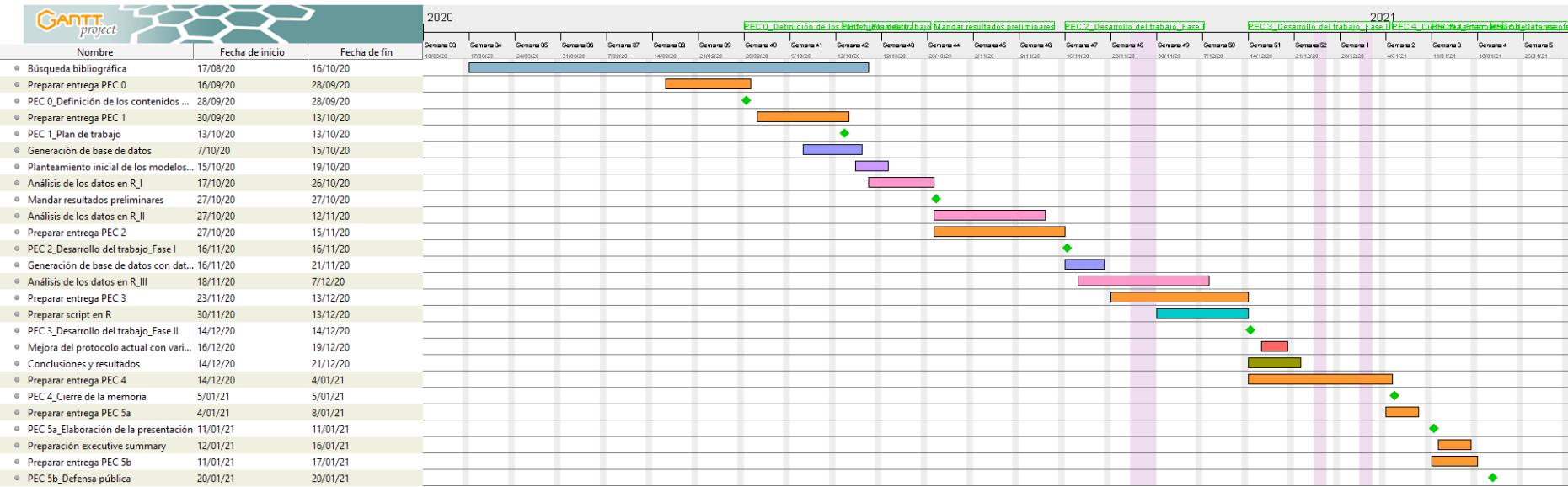


Figura 1: cronograma del proyecto

1.4.4 Hitos

Los hitos que se han mostrado mediante el Gantt en la imagen anterior se muestran también en la Tabla 2. Todos ellos coinciden con los que señala el plan docente de la asignatura para cada uno de los entregables.

Tabla 2: hitos para la elaboración del proyecto

Hito	Descripción	Fecha crítica
PEC 0	Definición de los contenidos del trabajo	08.09.2020
PEC 1	Plan de trabajo	13.10.2020
PEC 2	Desarrollo del trabajo, fase I	16.11.2020
PEC 3	Desarrollo del trabajo, fase II	14.12.2020
PEC 4	Cierre de la memoria	05.01.2021
PEC 5a	Elaboración de la presentación	11.01.2021
PEC 5b	Defensa pública	20.01.2021

1.4.5 Análisis de riesgos

En este subapartado se plasma el análisis de riesgos definido al comienzo del proyecto, ya que existen algunos factores que se cree que pueden repercutir de forma negativa en el desarrollo del mismo. Estos riesgos se muestran a continuación:

1. Falta de medición de uno de los biomarcadores (cortisol) en las muestras biológicas de saliva recogidas en las visitas a los hogares para el estudio piloto comenzado en la Universidad de Maryland. A día de hoy, en el laboratorio IDC Herzliya (Israel), donde fueron enviadas las muestras y actualmente éstas se encuentran, únicamente se ha analizado el biomarcador oxitocina. Debido a la pandemia del SARS-CoV-2/COVID-19, el laboratorio está priorizando su actividad a muestras relacionadas con el virus, por lo que no se sabe si las muestras serán analizadas durante el presente semestre.
En el caso que las muestras no fueran medidas, la evaluación se haría de forma teórica para ese biomarcador.
2. El resultado de significancia de la variable etnia podría ser otro factor de riesgo. Podría ocurrir que una vez definido el modelo con la variable explicativa etnia añadida en él, ésta no sea significativa. Esto daría lugar a una conclusión. Sin embargo, hay que tener en cuenta que el número de participantes en el estudio piloto comenzado en la Universidad de Maryland es muy reducido, y que la diversidad étnica de los individuos no es grande en este momento. Tal y como se ha definido en los objetivos, existe la intención de seguir con el estudio mediante un proyecto conjunto entre la UMD y el NIH, y de este modo, al incluir a más individuos en el estudio la base de datos será más grande. Al ejecutar el script generado en R con la nueva base de datos, la significancia de la variable etnia podría cambiar y los resultados podrían ser diferentes a los obtenidos cuando finalice el presente proyecto.
3. Los modelos planteados utilizando datos previamente publicados incluyan covariables significativas que no se habían considerado durante la recogida de datos del proyecto piloto. Esto impediría aplicar los modelos sobre los datos.

1.5 Breve sumario de productos obtenidos

Al final de este trabajo de fin de máster, se espera obtener una memoria que describa detalladamente el trabajo realizado a lo largo del semestre, para que posteriormente se presente de forma virtual ante un tribunal y de este modo se completen los estudios del Máster en Bioinformática y Bioestadística.

Los productos que se esperan conseguir son modelos ideales para cada uno de los biomarcadores relacionados con el estrés y la capacidad de generar vínculos sociales, que mediante variables explicativas muestren el efecto que éstas tienen sobre ellos. Estas variables serán en su gran mayoría factores económicos y demográficos.

Además de los modelos, dado el pequeño número de participantes en el estudio piloto comenzado en 2018 y como se espera que se pueda seguir con la recogida de datos en posteriores visitas a hogares con más sujetos involucrados, se quiere generar un script en R, al cual únicamente haya que introducirle la base de datos que se genere y que automáticamente se obtenga el efecto de la etnia sobre los marcadores biológicos. Esto hará que el modelo que se plantee en este trabajo tenga una sensibilidad mayor.

Asimismo, al finalizar el trabajo se espera hacer un resumen (*executive summary*) para poder llevar a cabo una presentación en el departamento correspondiente de la Universidad de Maryland que resuma los aspectos más relevantes del proyecto.

Finalmente, dado el corto periodo de tiempo para llevar a cabo el proyecto, no es posible que los productos adicionales que puedan salir de este trabajo estén publicados para enero 2021 (fecha en la que se presenta el presente trabajo). Estos productos a medio-largo plazo son los siguientes:

- Elaboración de una propuesta de proyecto entre la Universidad de Maryland y el *National Institute of Health* para continuar con la investigación en el *Department of Human Development and Quantitative Methodology* liderado por la Dr. Cabrera y en colaboración con la Dr. Feldman en IDC Herzliya (Israel).
- Elaboración de un artículo en una revista científica.

1.6 Breve descripción de los otros capítulos de la memoria

En esta sección se describe brevemente lo que se presentará en los siguientes capítulos de la memoria. El segundo apartado (Metodología), recoge la parte central y más extensa del trabajo, detallando cada paso realizado para obtener los modelos que describen los biomarcadores de la oxitocina y el cortisol. Previo al desarrollo del modelo de cada biomarcador, se lleva a cabo una descriptiva general de las variables incluidas en el conjunto de datos utilizado (sección 2.2). El apartado 2.3 recoge el análisis del biomarcador de la oxitocina, pero no es hasta el apartado 2.3.5 donde se describe el modelo, ya que previamente se realiza un análisis más concreto de las variables, incluyendo gráficas y tablas descriptivas. El subapartado del modelo describe únicamente el modelo con mejores resultados, añadiendo en los anexos de la memoria los modelos descartados. Además del modelo y la ecuación final que lo describe, también se realiza un análisis del comportamiento de los residuos, tanto de manera gráfica como aplicando diferentes tests. El análisis del cortisol se realiza siguiendo el mismo formato que para la oxitocina, pero en este caso se analizan tres modelos diferentes: 1) utilizando el conjunto de datos completo 2) utilizando únicamente las mediciones de la sangre y 3) utilizando las mediciones en la saliva. En el apartado 2.5 se describe la aplicabilidad de los modelos de forma teórica. El apartado 2.6 describe brevemente el repositorio en el que se ha hecho público el trabajo realizado. En el tercer apartado del documento se describen las conclusiones obtenidas tras el desarrollo del presente trabajo.

2. Metodología

En este apartado se describe el proceso para el desarrollo de los modelos utilizando datos de la literatura. El apartado está dividido en diferentes secciones, que se describen a continuación.

2.1. Planteamiento inicial y generación de la base de datos

La posibilidad de utilizar una base de datos generada a partir de las visitas a los hogares de familias en Estados Unidos (a través de la Universidad de Maryland) quedó descartada debido a la pandemia actual del SARS-CoV-2/COVID-19. Las medidas de distanciamiento social han impedido realizar las visitas, y por lo tanto, no se han podido recopilar más datos sociodemográficos y biológicos de nuevos participantes. Durante el periodo previo a la pandemia, se habían realizado nueve visitas a hogares, y por lo tanto se había recogido información de padres, madres y bebés de nueve núcleos familiares distintos. Al tratarse de una muestra muy reducida para plantear los modelos asociados a cada biomarcador, se ha llevado a cabo un estado del arte con el objetivo de encontrar artículos previamente publicados donde se analice uno o ambos biomarcadores y los relacione de alguna manera con el estrés. En la Tabla I.A del Anexo A se pueden observar los artículos seleccionados y la fecha en la que se contactó al autor/a correspondiente de los estudios mediante correo electrónico. Como se puede observar en la tabla, hubo un porcentaje de respuestas muy reducido, ya que de los 29 artículos seleccionados y contactados, únicamente 7 investigadores respondieron (24,14%), de los cuales N=2 (6.89%) contestaron que se debía contactar a otros co-autores de los estudios (a los que se contactó pero no se obtuvo respuesta), N=2 (6.89%) no tenían los derechos de sus instituciones para compartir el conjunto de datos o estaban todavía trabajando con ellos, N=2 (6.89%) únicamente mandaron datos agregados a partir de los resultados obtenidos ya que no se les permitía compartir el conjunto de datos, y un único autor (N=1, 3.45%) envió el conjunto de datos para su uso en este estudio académico. El autor, sin embargo, no pudo compartir los datos asociados a las mediciones del ritmo cardíaco que había medido y utilizado durante el estudio debido a la negativa por parte de su institución. La segunda base de datos utilizada para completar el conjunto de datos estaba disponible para su descarga.

La base de datos que se ha utilizado para definir los modelos de la oxitocina y el cortisol tras someter a los individuos a situaciones de estrés se han obtenido a partir de los artículos de Tas et al. 2018 y Ooishi et al. 2017. Ambos estudios analizan los cambios en los biomarcadores cortisol y oxitocina tras someter a los participantes a una situación de estrés. Para generar una única base de datos que unifique las observaciones y variables recogidas en ambos artículos, se generó un archivo Excel y posteriormente se ha cargado en el software R.

En total, la muestra está compuesta por 84 observaciones y 23 variables. De todas las observaciones, 32 son del artículo de Tas et al. 2018 y el resto pertenecen al estudio de Ooishi et al. 2017. Al tratarse de estudios totalmente independientes entre sí, no todas las variables están recogidas en ambos estudios, por lo que existe un porcentaje elevado de valores faltantes (NA) en algunas de las variables, las cuales se analizarán más adelante.

2.2. Descriptiva de los datos

Para conocer cada una de las variables que componen el conjunto de datos, a continuación se muestra la siguiente Tabla 3 descriptiva, que muestra el nombre de cada variable, el tipo de variable, el número de observaciones, los niveles existentes para las variables categóricas, los valores faltantes de la variable y una breve descripción de cada una de ellas.

Tabla 3: descriptiva datos base de datos

Nombre variable	Tipo de variable	Descripción	N	Niveles	Valores faltantes
id	Categórica	Variable identificativa para cada participante	84	58	0
age	Numérica	Edad de los participantes	84	-	0
gender	Categórica	Sexo de cada participante	84	2	0
disease	Categórica	Enfermedad diagnosticada	84	2	0
med.type	Categórica	Tipo de medicación	84	2	0
med.dos	Numérica	Dosis de la medicación (mg)	32	-	52
oral.count	Categórica	Ingesta de anticonceptivos orales	84	2	0
stimulus.type	Categórica	Tipo de estímulo utilizado para generar estrés en el estudio	84	3	0
co.meas	Categórica	Tipo de muestra cortisol	84	2	0
oxt.meas	Categórica	Tipo de muestra oxitocina	84	2	0
co.pre	Numérica	Nivel de cortisol antes del estímulo (pg/ml)	83	-	1
co.post	Numérica	Nivel de cortisol tras el estímulo (pg/ml)	84	-	0
oxt.pre	Numérica	Nivel de oxitocina antes del estímulo (pg/ml)	78	-	6
oxt.post	Numérica	Nivel de oxitocina tras el estímulo (pg/ml)	46	-	38
hr.bas	Numérica	Media del ritmo cardiaco antes del estímulo	52	-	32
hr.post	Numérica	Media del ritmo cardiaco tras el estímulo	52	-	32
arousal_level	Numérica	Nivel de excitación	52	-	32
valence_level	Numérica	Valencia. Criterio utilizado para medir la emoción	52	-	32
co.reac	Numérica	Índice de reacción al cortisol (%)	32	-	52
co.res	Categórica	Reacción frente a las alteraciones en el cortisol	32	2	52
PANSS_positive	Numérica	Media de los valores obtenidos para medir la serenidad de los síntomas positivos	32	-	52
PANSS_negative	Numérica	Media de los valores obtenidos para medir la serenidad de los síntomas negativos	32	-	52
PANSS_general	Numérica	Media de los valores obtenidos para medir la serenidad general de los síntomas	32	-	52

La Tabla 4 describe los niveles de las variables categoricas descritas en la tabla anterior (Tabla 3).

Tabla 4: niveles de las variables categóricas

Nombre variable	Tipo de variable	Niveles
id	Ordinal	58; 1-32 únicos; 33-84 (26 participantes únicos; se repiten)
gender	Binaria simétrica	2; 1=mujer, 2=hombre
disease	Binaria asimétrica	2; 0= ninguna, 1=esquizofrenia
med.type	Binaria asimétrica	2; 0= ninguna, 1=CPZ (Chlorpromazine, mg)
oral.count	Binaria asimétrica	2; 0= no, 1=sí (solo aplicable a mujeres)
stimulus.type	Nominal	3; 0=ninguno , 1=psicológico, 2=físico
co.meas	Binaria asimétrica	2; 1=saliva, 2=sangre
oxt.meas	Binaria asimétrica	2; 1=saliva, 2=sangre
co.res	Binaria asimétrica	2; 1=no-respondedor, 2=respondedor

La base de datos está compuesta por observaciones de dos estudios totalmente independientes, y es por ello por lo que algunas de las variables no son comunes en ambos casos, generando una proporción elevada de valores NA en algunas variables que componen la base de datos, tal y como se ha mostrado en la Tabla 3. Esto ocurre con las variables *PANSS_*, *oxt.post*, *hr.bas*, *hr.post*, *arousal_level* y *valence_level*, que únicamente se han utilizado en uno de los dos estudios (Ooishi et al. 2017). Sin embargo, el uso de las demás variables (u observaciones completas) son suficientes para generar diferentes modelos estadísticos.

En ambos artículos han utilizado diferentes métodos para medir el nivel de los biomarcadores: en el estudio de Tas et al. 2018 ambos biomarcadores se miden en la sangre (*serum level cortisol*) y en el estudio de Ooishi et al. 2017 en la saliva. El artículo de Kaufman et al. 2002 muestra que las concentraciones de las hormonas en saliva son más bajas que las de la sangre. Sin embargo, esto no es un problema si para esa hormona, la medida en saliva está correlacionada con la muestra recogida en la sangre. En el caso de la oxitocina, el artículo de McCullough, Churchland, y Mendez 2013 muestra que ambas muestras están relacionadas en un 50%, y en el caso del cortisol, la relación es más alta, hasta llegar a una relación del 90%, tal y como demuestra el artículo de Peters et al. 1982. En el caso del cortisol, hay que tener en cuenta que no es lo mismo el nivel de cortisol general o el cortisol libre, y que la proporción de correlación entre saliva y sangre no se debe aplicar en estos casos puesto que se trata de medidas diferentes.

La diferencia más significativa entre ambos artículos se observa en las medidas del cortisol, debido a las siguientes razones: 1) la medida en sangre mide el cortisol general, y la medida en saliva mide el nivel de cortisol libre y 2) las unidades en las que se ha medido el cortisol en cada artículo es diferente. Para llevar a cabo el análisis, primero se han transformado las unidades del cortisol en sangre para que estén en las mismas unidades que en la saliva (transformar de $\mu\text{g}/\text{dl}$ a pg/ml). Posteriormente, se ha transformado el nivel de cortisol existente en la sangre en cortisol libre para que se iguale al de la saliva. Los artículos de Estrada-Y-Martin y Orlander 2011 y Hammond, Smith, y Underhill 1991 afirman que entre el 80% y el 90% del cortisol en sangre está unido a CBG - *Cortisol Binding Globulin*, que el 5% y el 10% está unido a la albumina, y que por lo tanto, como máximo únicamente el 5% del cortisol en sangre es cortisol libre. Tanto los valores previos del cortisol al estímulo de estrés como los posteriores se han multiplicado por 0.025, para que únicamente se tuviera en cuenta la cantidad de cortisol libre y así poder compararlo con los valores en la saliva. En el caso de los valores medidos para la oxitocina, éstos no han requerido de ninguna transformación entre ambos conjuntos de datos puesto que

ambos se han medido originalmente en la misma unidad (pg/ml) y la diferencia entre la sangre y la saliva no se ha considerado un problema.

El artículo de Ooishi et al. 2017 es el único que ha medido los valores del biomarcador oxitocina tras la aplicación del estímulo en el participante. Para el correcto desarrollo del trabajo, y puesto que el objetivo es generar un modelo para cada biomarcador, el conjunto de datos se ha dividido en dos, recogiendo en cada uno de ellos los datos de oxitocina y cortisol respectivamente. El proceso para cada uno de ellos se muestra en las siguientes secciones.

2.3. Biomarcador I: Oxitocina

Para llevar a cabo el modelo que prediga el nivel de oxitocina tras someter a una persona a un estímulo, lo primero que se ha hecho ha sido separar la base de datos principal y eliminar aquellas variables relacionadas con el cortisol utilizando la función *select* del paquete *dplyr* ya que el objetivo no es ver cómo la variable respuesta (la oxitocina en este caso) cambia respecto a otro biomarcador, si no ver cómo varía en función de las variables demográficas y sociales descritas en la Tabla 3.

La base de datos generada para el análisis de la oxitocina se denomina *data.oxt* y está compuesta en un principio por 84 observaciones y 13 variables, que son las siguientes: *id*, *age*, *gender*, *disease*, *med.type*, *med.dos*, *oral.count*, *stimulus.type*, *oxt.meas*, *oxt.pre*, *oxt.post*, *hr.bas* y *hr.post* (explicadas y descritas en la Tabla 3). Sin embargo, es necesario realizar un análisis de los datos para observar el comportamiento de las variables y ver si es necesario mantener todas ellas en el conjunto de datos. Posteriormente, se planteará el modelo sobre las variables de interés.

2.3.1 Variable respuesta

La variable respuesta del modelo que se planteará en las siguientes secciones es *oxt.post*, que analiza el nivel de oxitocina tras aplicar el estímulo sobre el participante. Esta variable se ha definido en la Tabla 3 y se trata de una variable cuantitativa continua. Para obtener una descriptiva general de la variable, en la Figura 2 se muestra un gráfico de cajas de esta variable:

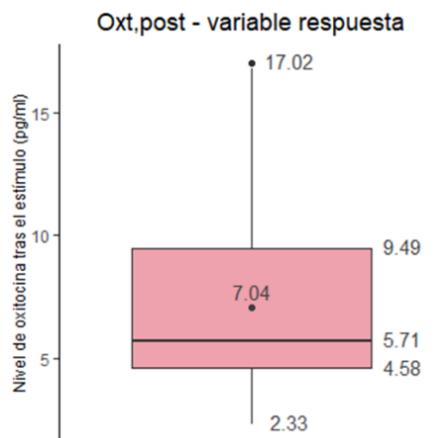


Figura 2: boxplot de la variable oxitocina tras aplicar un estímulo sobre el participante, donde se muestran los valores de la media, mediana Q1, Q3, min y max

En la Tabla 5 se muestran los valores más significativos de la variable respuesta *oxt.post* (el valor mínimo, máximo, la mediana, la media -junto con la desviación estándar- los cuantiles Q1 y Q3, así como los valores de las medidas de dispersión-varianza, Rango e IRQ-). La media de los participantes es de 7.04 pg/ml, con una desviación estándar de 3.77. En el gráfico se puede observar también un valor *outlier*, que hace referencia al valor máximo de la variable en el conjunto de datos, igualado a 17.02 pg/ml.

Tabla 5: descriptiva numérica de la variable respuesta oxt.post (nivel de oxitocina tras aplicar un estímulo sobre el participante)

	Oxt.post
Valor general	
Min	2.33
Q1	4.58
Mediana	5.71
Media (SD)	7.04 (3.77)
Varianza	14.22
Q3	9.49
Max	17.02
Rango	14.69
IQR	4.90

Aunque en el gráfico mostrado en la Figura 2 se observe la distribución de la variable, es necesario analizar si la variable cumple el supuesto de normalidad. Mediante la función *describe* del paquete *dlookr*, se obtiene que el valor que mide si existe simetría en la distribución de la variable (denominado *skewness*) es 1.04. Los valores cercanos a cero para la observación de *skewness* se pueden considerar simétricos, y cuanto mayor sea el valor obtenido en la observación, significará que la variable difiere más de una distribución normal. En este caso, la variable respuesta no se aleja demasiado del valor nulo, pero en la Figura 2 se ha intuido que la variable puede estar sesgada a la derecha, debido a la distribución observada en el tercer cuantil del análisis. El valor de *kurtosis*, analiza el grado de presencia de valores outliers en la distribución, y en este caso, se obtiene un valor menor que para el caso de *skewness*, por lo que no parece que los valores *outliers* vayan a suponer un problema durante el análisis.

Es importante analizar utilizando diferentes test si la variable sigue una distribución normal. En este caso se ha analizado mediante el test de *Shapiro-Wilk*, fijando el nivel de significancia en un 5% y analizando el p-valor obtenido para aceptar o no la hipótesis nula. Este test establece como hipótesis nula la existencia de una distribución normal de los datos, y para la hipótesis alternativa, la distribución no normal de los datos. Se aplica la función *normality* del paquete *dlookr*, y se obtiene un p-valor inferior al 5%, por lo tanto no se acepta la hipótesis nula y no se considera que la variable respuesta que mide la oxitocina post estímulo (*oxt.post*) siga una distribución normal. Para poder analizar gráficamente el comportamiento respecto a la normalidad, a continuación se muestra la Figura 3 con la distribución de la variable.

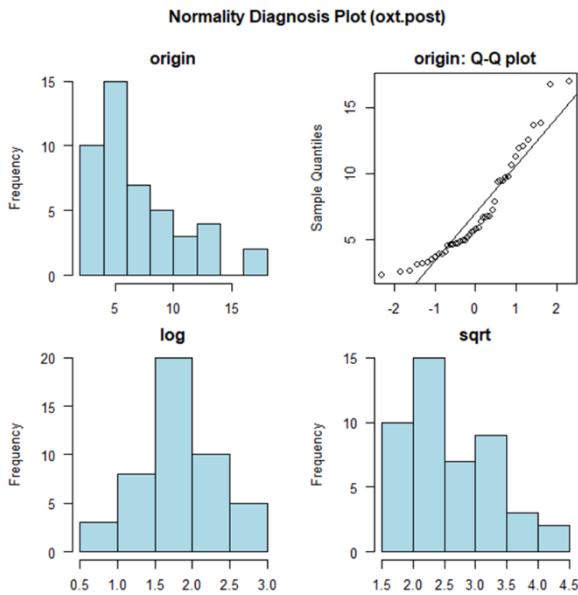


Figura 3: distribución de la variable respuesta que mide el nivel de oxitocina tras aplicar un estímulo sobre el participante (*oxt.post*). Arriba a la izquierda, histograma de la distribución original. Arriba a la derecha, gráfico QQ de los datos originales. Los gráficos de abajo muestran histogramas de la distribución de la variable en caso de aplicar la transformación logarítmica o de raíz cuadrada a los datos

En la figura anterior (Figura 3) se muestran cuatro gráficos: En el gráfico superior izquierdo, se muestra la distribución de la variable respuesta sin aplicar ninguna transformación sobre ella. Efectivamente, tal y como se preevía al observar el gráfico de cajas, la variable está sesgada a la derecha. En el gráfico superior derecho, también es posible observar cómo los puntos de cada una de las observaciones no se sobreponen con la línea que marca la normal. En los gráficos inferiores, se muestran dos planteamientos para transformar la variable respuesta: a la izquierda, la transformación logarítmica de la variable, donde se observa a simple vista que la variable podría estar distribuida de forma normal en caso de aplicar la transformación sobre ella, y a la derecha, la transformación de la raíz cuadrada de los datos. Por lo tanto, para comprobar que la transformación logarítmica asemeja la variable a una distribución normal, se aplica una vez más la función *normality* sobre ella, donde en este caso se obtiene un p-valor igual a 0.39, y por lo tanto no habría evidencia suficiente para rechazar la hipótesis nula del test de Shapiro-Wilk y se aceptaría la distribución normal de la variable respuesta que mide el nivel de oxitocina tras el estímulo (*oxt.post*).

2.3.2 Valores faltantes en el conjunto de datos

El conjunto de datos *data.oxt* está compuesto por 13 variables (incluyendo la variable respuesta *oxt.post* analizada previamente) y 84 observaciones. Sin embargo, no todas las variables serán adecuadas para predecir la variable respuesta de la oxitocina, puesto que algunas presentan muchos valores faltantes (NA) en sus observaciones. Además, la propia variable respuesta *oxt.post* tiene un porcentaje elevado de Nas. Se considera necesario analizar en detalle y ver en qué combinaciones y situaciones se observan los valores faltantes. Mediante la función *aggr* del paquete VIM, se visualiza en la Figura 4 la proporción de valores faltantes en el conjunto de datos.

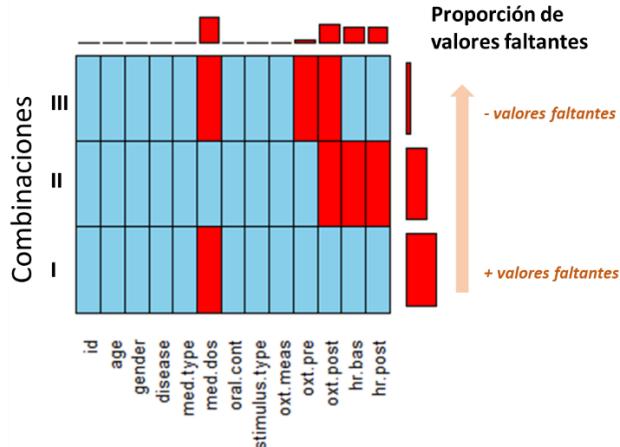


Figura 4: valores faltantes en el conjunto de datos data.oxt obtenido mediante la función aggr del paquete VIM.
Proporción de valores faltantes en variable numéricas (orden ascendente en la dirección marcada) para tres combinaciones

En la Figura 4 se observa que la variable que mide la dosis de medicación (*med.dos*) es la variable numérica que más valores faltantes incluye en el conjunto de datos (combinación I). Además, para la variable respuesta de la oxitocina (*oxt.post*), se observa que en los casos donde los valores de la variable *oxt.post* faltan, también lo hacen las mediciones del ritmo cardiaco (se trata de las observaciones referentes al artículo Tas et al. 2018, mostradas en la combinación II), y en los casos (menos frecuente) donde los valores de la oxitocina previa al estímulo (*oxt.pre*) faltan, también lo hacen los valores de *oxt.post* y *med.dos* (artículo Ooishi et al. 2017, mostrado en la combinación III). Se decide eliminar del conjunto de datos la variable *med.dos*, ya que representa el porcentaje más alto de valores faltantes en el conjunto de datos considerando todas las variables, con un 62% de valores faltantes.

Tras eliminar la variable de la dosis de medicamento (*med.dos*), en el conjunto de datos hay 12 variables y 84 observaciones. Sin embargo, los valores faltantes en la variable respuesta *oxt.post* pueden suponer un problema a la hora de generar el modelo, ya que se ha observado que de las 84 observaciones únicamente 46 están completas, 32 tienen valores faltantes en ambas variables del ritmo cardiaco (*hr.bas* y *hr.post*) y oxitocina post estímulo (*oxt.post*), y otras 6 observaciones tienen valores faltantes tanto en el nivel de oxitocina previo (*oxt.pre*) como en el posterior (*oxt.post*). Estos datos se observan de forma resumida en la Figura 5 que se muestra a continuación:

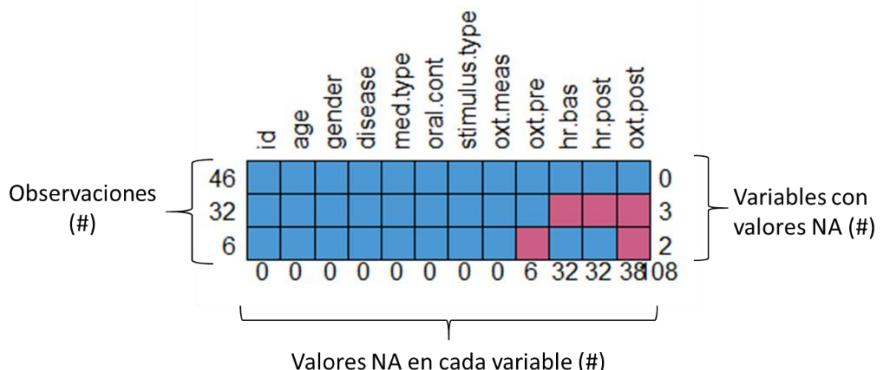


Figura 5: valores faltantes del conjunto de datos data.oxt eliminando la variable *med.dos*, donde hay 46 observaciones completas, 32 donde hay valores faltantes en tres variables y 6 observaciones con valores faltantes en dos variables

En la Tabla 6 que se muestra a continuación se resumen los valores mostrados en la imagen previa.

Tabla 6: número de observaciones completas e incompletas del conjunto de datos *data.oxt*, y descripción de cuales son las variables que tienen valores faltantes

Número de observaciones	Descripción
46	No falta ninguna observación
32	Valores faltantes en las variables <i>hr.bas</i> , <i>hr.post</i> y <i>oxt.post</i>
6	Valores faltantes en las variables <i>oxt.pre</i> y <i>oxt.post</i>

Como en 32 observaciones (38.1%) hay datos faltantes para la oxitocina post estímulo (*oxt.post*) y ésta es la variable respuesta de los modelos que se plantearán en las siguientes secciones, se decide eliminar las observaciones que no estén completas del conjunto de datos, manteniendo la variable en la base de datos. Para ello, se filtran las observaciones no completas del conjunto de datos *data.oxt* mediante la función *complete.cases()*. No se elimina la variable aunque tenga un porcentaje elevado de *missings* puesto que se trata de la variable dependiente que se usará en el modelo que se plantea en la sección 2.3.5. Antes de filtrar los datos, había 108 valores NA en total, y tras la eliminación de todos los valores faltantes, el conjunto de datos está compuesto por 46 observaciones y 12 variables.

La filtración de los datos y mantener únicamente las observaciones completas ha hecho que las variables binarias categóricas de género (*gender*), enfermedad (*disease*) y tipo de muestra recogida (*oxt.meas*) únicamente tengan un nivel de respuesta, por lo tanto no se incluirán en los modelos que se plantearán en las siguientes secciones, puesto que no permiten la comparación con otros niveles para esa misma variable. También se elimina la variable categórica que mide si se ingieren o no anticonceptivos orales (*oral.count*), puesto que ninguna participante de los estudios citados tomaba anticonceptivos orales y la variable no añade por lo tanto información al estudio.

Finalmente, el conjunto de datos que recoge las posibles variables que se deberían utilizar a la hora de diseñar un modelo para el biomarcador oxitocina, se compone de 46 observaciones y 6 variables.

2.3.3 Variables predictoras

De las 6 variables que componen el conjunto de datos, 5 se consideran variables predictoras, ya que la sexta es la variable respuesta. Estas variables son las siguientes: edad (*age*), tipo de estímulo aplicado para generar estrés en el participante (*stimulus.type*), nivel de oxitocina previo al estímulo (*oxt.pre*) y ritmo cardíaco antes y después del estímulo (*hr.bas* y *hr.post* respectivamente), todas ellas descritas en la Tabla 3. A excepción de la variable *stimulus.type*, las demás variables son cuantitativas. La variable *age*, es la única variable cuantitativa discreta, y las demás, son variables cuantitativas continuas. La variable *stimulus.type*, es una variable categórica con dos niveles para el análisis de la oxitocina: tiene el valor de 0 cuando no se aplica un estímulo estresante sobre la persona y coge el valor de 2 cuando el estímulo de estrés se aplica sobre el participante de manera física. Aunque la variable que mide el tipo de muestra de oxitocina analizada (*oxt.meas*, con los niveles de saliva o sangre) se haya eliminado del conjunto de datos final *data.oxt*, es importante destacar que todas las variables se han medido mediante muestras de saliva. Al haber únicamente una variable categórica en el conjunto de datos, no es

possible reportar tablas cruzadas entre las variables no-numéricas. Sin embargo, a modo de resumen cabe destacar que hay 23 observaciones donde no se aplica ningún estímulo (por lo tanto, *stimulus.type==0*) y otras 23 observaciones cuando *stimulus.type se iguala a 2*, es decir se aplica un estímulo físico. Tal y como se ha realizado para la variable respuesta, a continuación se muestra la distribución de las variables numéricas *oxt.pre* (oxitocina previa al estímulo), *hr.bas* y *hr.post* (ritmos cardiacos antes y después del estímulo respectivamente) según el tipo de estímulo aplicado sobre ellas:

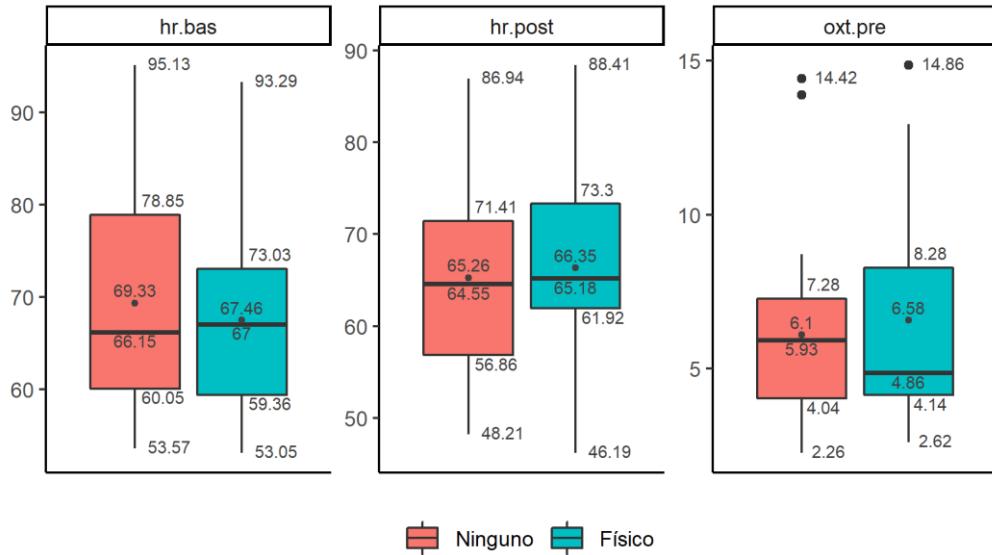


Figura 6: boxplots con los valores de la media, mediana, Q1, Q3, min y max para las variables del ritmo cardiaco (*hr.bas*-izq.-, *hr.post* –centro-) y oxitocina previa al estímulo (*oxt.pre*-dch.-) separados según si no se aplica ningún estímulo o se aplica un estímulo físico sobre el participante

A simple vista, en ninguno de los tres gráficos de la Figura 6 se observa que la variable esté distribuida de forma normal. En algunos grupos (*hr.bas* sin estímulo u *oxt.pre* con y sin estímulo) parece que las variables están muy sesgadas debido al tamaño de los cuantiles en cada caso. Para analizar los valores numéricamente, en la Tabla 7 que se muestra a continuación se describen los valores de las tres variables mostradas en la figura previa, tanto de forma general como clasificándolas por cada tipo de estímulo aplicado.

Tabla 7: descriptiva numérica de las variables *oxt.pre*, *hr.bas* y *hr.post*, tanto de forma general como separandolas por el tipo de estímulo aplicado sobre ellas. Se recogen valores generales (min, max, media, mediana, Q1, Q3) y valores de las medidas de dispersión de cada una (varianza, rango, IQR)

	Variable		
	Oxt.pre	Hr.bas	Hr.post
Valor general			
<i>Min</i>	2.26	53.05	46.19
<i>Q1</i>	4.07	60.04	58.99
<i>Mediana</i>	5.39	66.47	64.79
<i>Media (SD)</i>	6.34 (3.33)	68.4 (10.72)	65.8 (9.91)
<i>Varianza</i>	11.12	114.98	98.25
<i>Q3</i>	7.3	75.63	72.42
<i>Max</i>	14.86	95.13	88.41
<i>Rango</i>	12.6	42.08	42.62
<i>IQR</i>	3.22	15.59	13.43
Ningún estímulo			
<i>Min</i>	2.26	53.57	48.21
<i>Q1</i>	4.03	60.04	56.85
<i>Mediana</i>	5.93	66.15	64.55
<i>Media (SD)</i>	6.1 (3.08)	69.33 (11.18)	65.26 (10.23)
<i>Varianza</i>	9.48	125.0	104.66
<i>Q3</i>	7.28	78.855	71.41
<i>Max</i>	14.42	95.13	86.94
<i>Rango</i>	12.16	41.56	38.73
<i>IQR</i>	3.24	18.81	14.55
Estímulo físico			
<i>Min</i>	2.62	53.05	46.19
<i>Q1</i>	4.14	59.36	61.92
<i>Mediana</i>	4.86	67.00	65.18
<i>Media (SD)</i>	6.58 (3.63)	67.46 (10.41)	66.34 (9.78)
<i>Varianza</i>	13.15	108.37	95.70
<i>Q3</i>	8.27	73.03	73.3
<i>Max</i>	14.86	93.29	88.41
<i>Rango</i>	12.24	40.24	42.22
<i>IQR</i>	4.13	13.66	11.38

En la distribución que se presenta en la Figura 7 mediante el gráfico de cajas de la variable numérica *age*, se muestran todas las observaciones en un mismo grupo, puesto que de las 46 observaciones del conjunto de datos *data.oxt*, únicamente hay 23 pacientes que son únicos. Es decir, ambos tipos de estímulos se han aplicado sobre los mismos participantes el mismo día (o días seguidos) y por lo tanto la distribución de la edad es la misma para ambos estímulos.

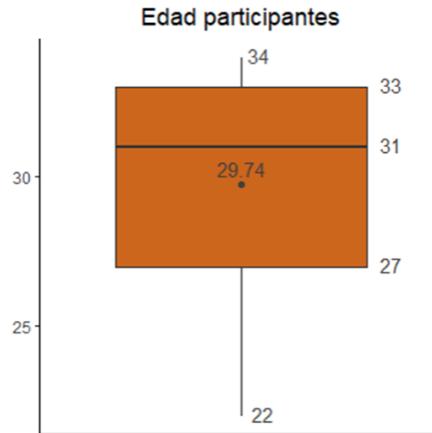


Figura 7: boxplot de la variable edad para ambos grupos de estímulos del conjunto de datos data.oxt. Se muestran los datos numéricos del mínimo, máximo, media, mediana, Q1 y Q3

De la misma manera que con las demás variables numéricas del conjunto de datos de la oxitocina, en la Tabla 8 se muestran los valores de la variable edad con las medidas de dispersión incluidas.

Tabla 8: descriptiva numérica variable edad, donde se recoge el valor mínimo, máximo, cuantiles, media, y valores de las medidas de dispersión (varianza, rango, IQR)

	Age
Valor general	
<i>Min</i>	22.00
<i>Q1</i>	27.00
<i>Mediana</i>	31.00
<i>Media (SD)</i>	29.74 (3.89)
<i>Varianza</i>	15.13
<i>Q3</i>	33.00
<i>Max</i>	34.00
<i>Rango</i>	12.00
<i>IQR</i>	6.00

Para analizar el comportamiento general de las variables, es posible observar el valor de *skewness* para la simetría y el valor de *kurtosis* para los valores *outliers* de las variables numéricas. En este caso, la variable cuyo valor de *skewness* es más alto es *oxt.pre*, con un valor de 1.22, muy parecido al obtenido para la variable respuesta.

Aunque a simple vista y en base a los valores de *skewness* obtenidos mediante la función *describe*, ninguna de las variables numéricas sigue una distribución simétrica, por lo tanto no cumpliría con la hipótesis de la normalidad. Para ello, se aplica la función *normality()* sobre los datos, que mide mediante el test de Shapiro-Wilk si la variable está distribuida de forma normal, fijando el nivel de significancia en un 5%. Del test se obtiene que la variable que menos se asemeja a una distribución normal es la que mide el nivel de oxitocina previo al estímulo (*oxt.pre*) con un p-valor de $5.99 \cdot 10^{-5}$, seguida de la variable edad. En las únicas variables donde no existe evidencia suficiente para rechazar la hipótesis nula debido a que obtiene un p-valor superior al 5% son ambos ritmos cardiacos (*hr.post* y *hr.bas*). Es aconsejable analizar la distribución de las variables de forma gráfica para ver cómo se comportan y para ello a continuación se muestran los gráficos obtenidos a partir de la función *plot_normality* para las variables *oxt.pre*, *age*, *hr.bas* y *hr.post*.

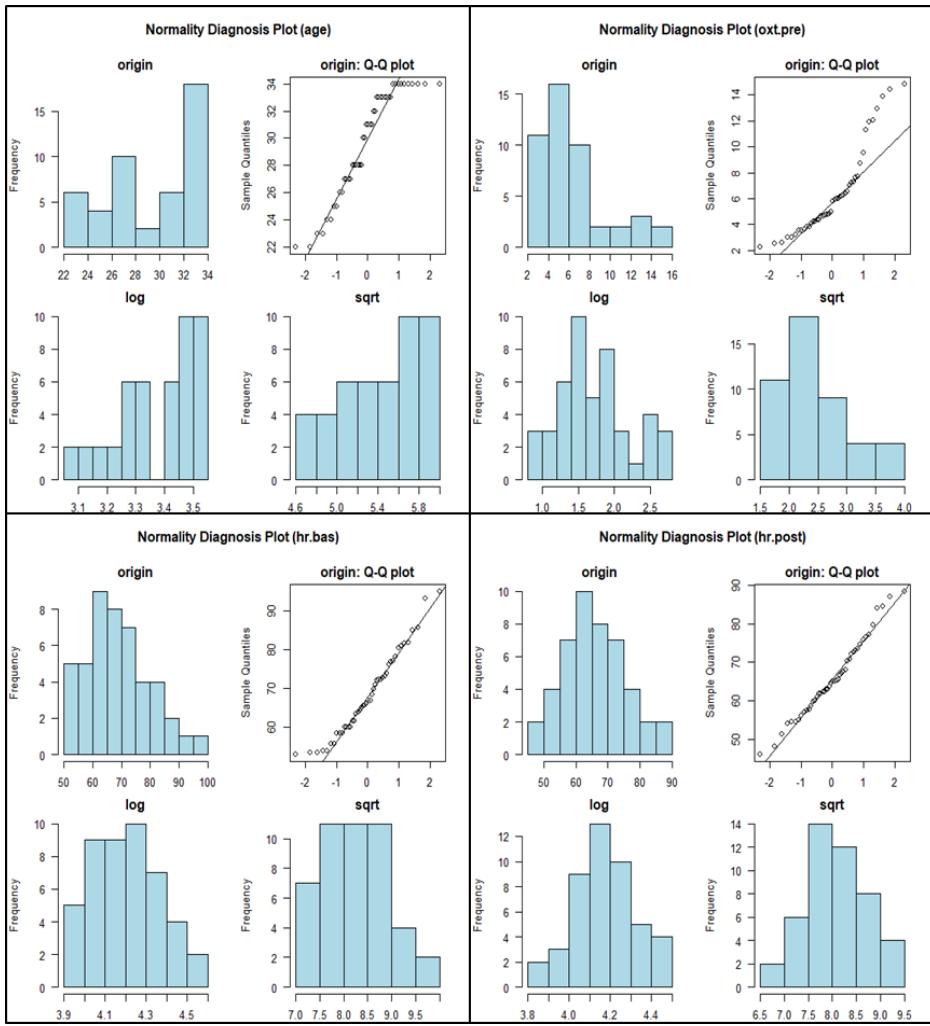


Figura 8: análisis de la normalidad para la variable edad (arriba izquierda), oxitocina pre-estímulo (oxt.pre, arriba a la derecha), ritmo cardiaco previo al estímulo (hr.bas, abajo a la izquierda) y ritmo cardiaco post estímulo (hr.post, abajo a la derecha). Para cada una se muestra la distribución original mediante histograma y gráfico QQ, e histograma con transformación log y sqrt

Los outputs de la función `plot_normality` en la Figura 8 para cada una de las variables numéricas mencionadas demuestra que el resultado que se observa está relacionado con el p-valor analizado, ya que el histograma en el que se observa una distribución normal sin aplicar ninguna transformación es el de la variable del ritmo cardiaco post estímulo `hr.post` ($p\text{-valor}=0.5$). En la variable del ritmo cardiaco previo `hr.bas` ($p\text{-valor}=0.08$) se observa que ésta podría estar sesgada a la derecha, y la transformación logarítmica simula una pequeña mejoría de la variable respecto a la original. En la variable que mide el nivel de oxitocina previo al estímulo (`oxt.pre`) se observa que ésta no está distribuida de forma normal, y que la distribución puede que mejore ligeramente al transformar logarítmicamente la variable. Finalmente, la variable edad muestra falta de normalidad a simple vista, tanto en la versión transformada como en la que no lo está. Si la variable edad se transformase logarítmicamente, el p-valor de la variable seguiría siendo muy pequeño ($p\text{-valor}=0.0002$) aunque de todos modos sería superior al p-valor obtenido sin aplicar la transformación. Para las variables `oxt.pre` y `hr.bas` transformadas logarítmicamente, sí que se obtiene un p-valor superior al 5% ($p\text{-valor } 0.22$ y 0.28 respectivamente), y por lo tanto no hay evidencia suficiente para rechazar la hipótesis nula en esos casos. Además, si se aplica la transformación logarítmica sobre la variable `hr.post`, aunque ya se aceptase la hipótesis nula de normalidad en su versión original, el valor del p-valor aumenta de 0.5 a 0.94 , por lo tanto se considera que mejora la normalidad de forma considerable.

2.3.4 Análisis de la correlación de variables

Para llevar a cabo el análisis de la correlación de las variables, y observar si existen correlaciones lineales entre la variable respuesta y las variables predictoras, se aplica la función *cor* sobre el conjunto de datos final. En la distribución de las variables analizada previamente se ha observado que alguna de las variables, al transformarlas logarítmicamente, mejoran su distribución y se asemejan a una distribución normal. Se aplica el método de correlación *Spearman*, en lugar del método *Pearson*, ya que aplicando el método de *Spearman*, se evita que el coeficiente de correlación varíe en el caso en el que la variable sea transformada. En la siguiente Tabla 9, se muestra la matriz de los coeficientes de correlación obtenida entre las variables del conjunto de datos:

Tabla 9: coeficientes de correlación de las variables del conjunto de datos *data.oxt* aplicando el método de *Spearman*

	age	stimulus.type	oxt.pre	oxt.post	hr.bas	hr.post
age	1					
stimulus.type	0	1				
oxt.pre	0.296	0.025	1			
oxt.post	0.217	-0.146	0.885	1		
hr.bas	0.311	-0.084	0.019	0.217	1	
hr.post	0.377	0.093	0.054	0.141	0.877	1

Es deseable que la variable respuesta (*oxt.post*) esté relacionada con las variables predictoras que definirán el modelo. Sin embargo, no es deseable que las variables predictoras, las cuales deben ser independientes, estén altamente correlacionadas con alguna otra variable predictoría. En este caso, se observa que la variable que mide el nivel de oxitocina post estímulo (*oxt.post*) tiene un coeficiente de correlación cuyo valor es 0.885 (positivo y muy alto) con la variable que mide la oxitocina antes de la aplicación del estímulo *oxt.pre* (la correlación entre ambas se muestra en la Figura 9). Es la correlación más alta existente entre la variable respuesta y cualquiera de las variables predictoras, ya que las otras tienen coeficientes de correlación inferiores a 0.22.

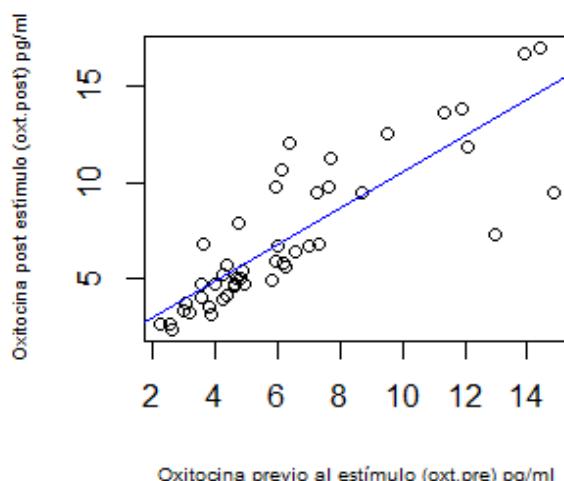


Figura 9: relación lineal entre la variable predictoria de oxitocina previa al estímulo (*oxt.pre*) y la variable respuesta de oxitocina post-estímulo (*oxt.post*)

Al analizar la correlación entre las variables predictoras, en la Tabla 9 se observa una correlación muy alta entre ambas variables que definen el ritmo cardiaco, *hr.bas* y *hr.post*, con un coeficiente de correlación igualado a 0.877. Esta correlación tan elevada supone que a la hora de plantear los modelos, una de ellas deba excluirse como variable predictora, para que los coeficientes que se obtengan en el modelo sean fiables y se evite la multicolinealidad en el modelo final. También es posible analizar la correlación entre las variables según el p-valor, y ver cuáles son significativos al 5%: en este caso, se obtiene un p-valor significativo para la combinación entre ambas variables del ritmo cardiaco ($p\text{-valor} = 2.22 \cdot 10^{-16}$), y también para la combinación de cada una de ellas con la variable edad (aunque con un p-valor más cercano a 0.05).

Para visualizar la correlación entre las variables del conjunto de datos *data.oxt* de forma gráfica, a continuación se muestra la Figura 10, donde los gráficos se han obtenido a través de la función *corrplot*:

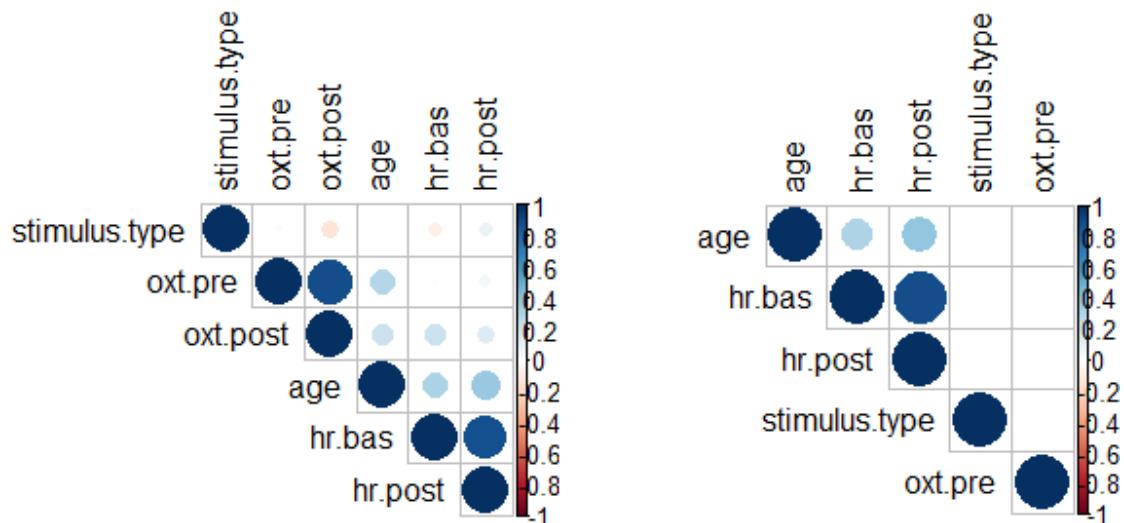


Figura 10: coeficientes de correlación del conjunto de datos *data.oxt* visualizados de forma gráfica. A la izquierda, todos los coeficientes, y a la derecha, visualización de los coeficientes significativos al 5%. Círculo más grande y oscuro, mayor correlación

En el gráfico de la izquierda de la figura anterior, se observa la correlación entre las diferentes variables predictoras y la variable respuesta. En este caso, cuanto más oscuro y grande sea el círculo, mayor correlación habrá entre las variables. En relación a las variables predictoras, se observa que los ritmos cardíacos están correlacionados, y en menor medida, la variable edad con ambas mediciones. También se observa correlación entre ambos niveles de oxitocina (*oxt.pre* y *oxt.post*). En el gráfico de la derecha, se muestran también los coeficientes de correlación pero eliminando aquellos valores de las variables predictoras que no son significativos al 5%. Una vez más, la mayor correlación se observa en la combinación de las medidas en los ritmos cardíacos y en la edad con ambas medidas. Finalmente, para concluir el análisis de la correlación, a continuación en la Figura 11 se muestra un mapa de calor (*heatmap*) con los valores de la matriz de correlación mostrada previamente.

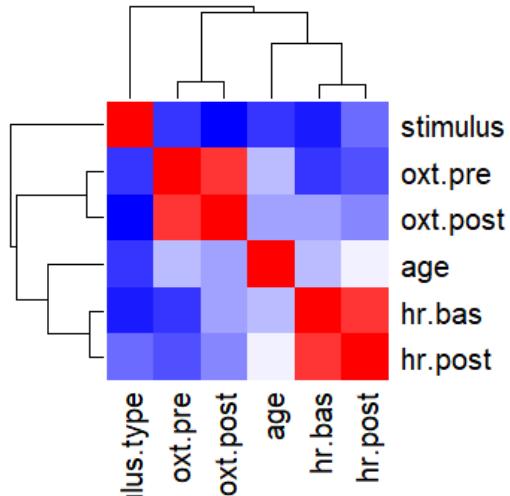


Figura 11: heatmap para el análisis de la correlación entre las variables del conjunto de datos *data.oxt*. Los rectángulos rojos identifican los coeficientes de correlación más cercanos a uno (más intensidad de rojo mayor correlación), y los rectángulos azules, menor correlación (mayor intensidad de azul menor correlación)

En el mapa de calor (*heatmap*) se observa que la correlación entre los ritmos cardiacos es muy alta, tal y como se ha ido observando desde el principio del análisis, y para la variable respuesta, ésta también muestra estar fuertemente correlacionada con la otra variable que mide el nivel de oxitocina (*oxt.pre*) tal y como se ha analizado durante el análisis.

2.3.5 Modelo

Una vez analizado el comportamiento de las variables en el conjunto de datos, en el presente subapartado se presenta el modelo con el que mejores resultados se han obtenido para predecir el valor de la variable respuesta *oxt.post*, que mide el nivel de oxitocina tras aplicar el estímulo sobre el participante. El modelo tiene que cumplir ciertas características, y una de ellas es la independencia de las variables predictoras. Sin embargo, de las 5 covariables, se ha observado que dos de ellas están altamente correlacionadas, por lo que no se pueden incluir ambas en el modelo que se plantea para evitar que se genere el principio de multicolinealidad. Para el análisis de la oxitocina, se plantea eliminar la covariable que mide el ritmo cardiaco post aplicación del estímulo (*hr.post*), puesto que muestra una menor correlación lineal con la variable dependiente (*oxt.post*) y además, el valor del R^2 es también inferior que el obtenido con el modelo que incluye únicamente el nivel del ritmo cardiaco previo al estímulo, *hr.bas* ($R^2 = 0.859$ frente al $R^2 = 0.52$ obtenido con el modelo donde se incluye la covariable *hr.post*). En el Anexo B, se incluye el desarrollo de otros modelos planteados, los cuales han sido finalmente descartados dado que el modelo que se presenta a continuación muestra mejores resultados, sobre todo respecto al comportamiento de los residuos del modelo. El primer modelo planteado en el anexo (sin ninguna transformación en los datos), se ha descartado debido a que no cumplía las suposiciones básicas de normalidad y homocedasticidad de los residuos. El segundo y el tercer modelo, donde en ambos casos se ha transformado la variable dependiente (transformación logarítmica y Box-Cox respectivamente) mostraba un peor comportamiento de los residuos respecto a la linealidad. Además, en los tres casos el valor del R^2 era inferior que el del presente modelo. Finalmente, en ninguno de los tres modelos descartados la variable edad era significativa al 5%, aunque tras aplicar el método de *stepwise selection*, en los tres casos ha resultado que se debía mantener pese a no ser significativa.

El modelo que se presenta se denomina *mod.oxt2*, donde todas las variables numéricas (tanto variable respuesta como predictoras) se han transformado logarítmicamente, aumentando el valor del R^2 ajustado y mejorando la distribución normal de los residuos. Además, tal y como se observa en el Anexo B, en los modelos donde no se ha aplicado ninguna transformación o

únicamente se ha aplicado la transformación logarítmica en la variable dependiente, los residuos de los modelos no se asemejan a la distribución normal, sobre todo en las colas de la distribución, donde muestran varios puntos *outliers*. Sin embargo, al aplicar la transformación logarítmica sobre todas las variables numéricas, se reduce la variabilidad de los residuos. Se trata por lo tanto de un modelo con transformación doble-log en las variables numéricas, añadiendo también una covariable categórica al modelo.

La fórmula del modelo que se plantea es la siguiente:

$$\log(Y) = B_0 + B_1 \log(X_{age}) + B_2 (X_{stimulus.type}) + B_3 \log(X_{oxt.pre}) + B_4 \log(X_{hr.bas}) + \epsilon$$

Ecuación 1: planteamiento inicial del modelo para predecir el nivel de oxitocina tras aplicar un estímulo sobre un participante con las covariables numéricas y la variable dependiente transformadas logarítmicamente

En el software R, el modelo se ha aplicado mediante la función *lm*, y el resultado que se obtiene del modelo se muestra en la Tabla 10 que se presenta a continuación:

Tabla 10: resultado del modelo de regresión mod.oxt2 para predecir el nivel de oxitocina post aplicación de un estímulo sobre un participante, con cuatro covariables: age, oxt.pre y hr.bas trasnsfromadas logarítmicamente y el tipo de estímulo

Predictores	Coeficiente B	Std.Err	t	Sig
Constante	-1.32512	0.92076	-1.439	0.157696
log(age)	-0.60697	0.23595	-2.572	0.013816 *
stimulus.type2	-0.16758	0.05731	-2.924	0.005604 **
log(oxt.pre)	1.00019	0.06243	16.022	< 2e-16 ***
los(hr.bas)	0.84390	0.20285	4.160	0.000158 ***
<i>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</i>				
F	69.56			
R ²	0.859			
p-valor	< 2.2e-16			

En la Tabla 10 se observa que el valor de R^2 ajustado es 0.859, y que todas las variables predictoras son significativas al 5%. Tras el planteamiento, es necesario analizar el comportamiento de los residuos del modelo, ya que en base a esos resultados, se podrá determinar si los coeficientes obtenidos para cada variable son fiables o no para estimar el valor de la variable respuesta. Analizar los residuos es importante, puesto que los errores del modelo lineal no deben seguir un patrón y de esta manera se evita poder predecir errores para las siguientes observaciones. A continuación, en la Figura 12, se muestran cuatro gráficos diferentes que describen los residuos del modelo *mod.oxt2*.

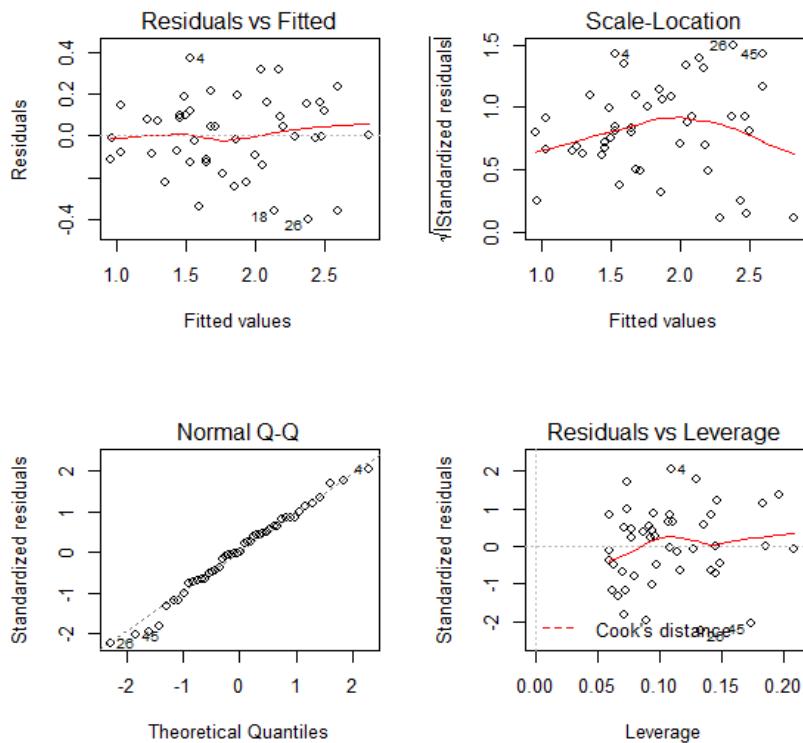


Figura 12: distribución de los residuos del modelo *mod.oxt2* (transformación doble log en las variables numéricas) para predecir el nivel de oxitocina tras aplicar un estímulo. Gráfico de linealidad (arriba izq.), homocedasticidad (arriba dcha.), normalidad (abajo izq.) y puntos outliers o influentes (abajo dcha.).

Cada uno de estos gráficos mostrados analiza diferentes aspectos de los residuos del modelo, los cuales se describen a continuación:

- Linealidad: analizado en el gráfico *Residuals vs Fitted*, que muestra si el modelo es una combinación lineal de las variables predictoras. Cuando los residuos son lineales, éstos se distribuyen alrededor de la línea horizontal. Para el modelo *mod.oxt2*, se observa que parece que este principio se cumple, ya que la línea roja está sobrepuerta en varios puntos a la línea horizontal central.
- Normalidad: analizado en el gráfico *Normal Q-Q*, que muestra si los residuos están distribuidos de forma normal. Para que se considere que los residuos están distribuidos de forma normal, éstos deberían estar encima de la línea discontinua. En este caso, observamos que las colas no están del todo alineadas con los valores centrales, pero parece que en general y a simple vista, la normalidad podría aceptarse ya que la mayoría de puntos están en el centro y éstos sí que se encuentran sobrepuertos.
- Homocedasticidad: analizado en el gráfico *Scale Location*, que muestra si la varianza de los residuos está distribuida de forma constante para las variables predictoras. En este caso se observa que la línea roja no es horizontal (por lo que puede ser que los residuos vayan cambiando para los valores predichos) y la distribución alrededor de la línea roja cuando los valores en el eje x (*fitted values*) aumentan parece que varían. El término contrario a la homocedasticidad es la heterocedasticidad, que sería el supuesto de que la varianza de los residuos no es constante, como parece ser el caso para el modelo *mod.oxt2*, aunque deberá de analizarse mediante el uso de diferentes tests.

- Detectar valores influentes (*outliers*) del modelo: mediante el gráfico *Residuals vs Leverage*. Los valores que se muestran separados del resto mediante la línea discontinua, son valores influentes, que de eliminarlos, el comportamiento del modelo podría cambiar. En este caso, se observa que hay algunos valores *outliers* (observaciones 4, 26 y 45) pero ninguno de ellos está separado por la distancia de Cook. Se ha descartado la posibilidad de eliminar los valores influentes del conjunto de datos para el planteamiento del modelo, ya que los residuos en caso de eliminarlos se comportan peor que los mostrados en la Figura 12, sobre todo en relación a la linealidad (es menos horizontal) y homocedasticidad (distribución más acampanada), aunque se sigan cumpliendo las suposiciones básicas para el modelo según los tests aplicados. Además, se ha analizado que de eliminarlos, aparecen nuevos valores influentes (en la primera ronda tras eliminar las observaciones numeradas, aparecen las observaciones 10, 17 y 34), y por lo tanto esto puede ser una indicación de que en lugar de valores *outliers*, la distribución de las variables del conjunto de datos está sesgada, y que siempre irán apareciendo más valores influentes cuando se eliminan los valores de las colas. Aunque en un principio pueda parecer que los valores mencionados podrían tener una gran influencia, finalmente en este caso se ha observado que no han sido casos extremos para predecir el valor de la variable respuesta y por lo tanto no han sido eliminados.

En resumen, a simple vista parece que el modelo es lineal y que los residuos están distribuidos de forma normal. Sin embargo, es necesario verificar estas suposiciones mediante diferentes tests sobre los residuos del modelo *mod.oxt2*.

2.3.5.1 Normalidad de los residuos:

Lo primero que se deberá hacer será verificar mediante un test de normalidad si los residuos del modelo *mod.oxt2* siguen o no una distribución normal, ya que gráficamente (en el gráfico Q-Q), podía observarse que las colas difieren de lo que se consideraría una distribución normal aunque a simple vista el resto sí que parece que cumple con la normalidad. Para comprobar la normalidad, se aplica la función *shapiro.test* del paquete *MASS* que hace referencia al test *Shapiro-Wilk*. Este test, asume en su hipótesis nula que los residuos siguen una distribución normal.

Tras aplicar el test sobre los residuos del modelo *mod.oxt2*, se obtiene un valor de $p=0.6364$, es decir, no existe evidencia suficiente para rechazar la hipótesis nula del test *Shapiro-Wilk* y por ello se asume que los residuos del modelo están distribuidos de forma normal, aunque en el gráfico en un principio haya parecido que la normalidad difería en las colas.

2.3.5.2 Homocedasticidad/heterocedasticidad:

Se analiza la homocedasticidad/heterocedasticidad del modelo *mod.oxt2* utilizando el test *Non-Constant Variance Score Test (ncVs)* y el test Breusch-Pagan. Ambos tests asumen en su hipótesis nula que la varianza de los residuos es constante (es decir, existe homocedasticidad) y en la hipótesis alternativa que la varianza cambia según los valores ajustados o la combinación lineal de las variables predictoras, es decir, existe heterocedasticidad.

En el modelo *mod.oxt2*, no hay evidencia suficiente para rechazar la hipótesis nula, ya que se obtiene un p-valor en cada test con valores de 0.14 y 0.59 respectivamente, y por ello se acepta que la varianza de los residuos es constante, y se asume que los residuos son homocedásticos. La existencia de homocedasticidad en los residuos del modelo se puede analizar también utilizando los tests de Levenne o Barlett, este último cuando se asume la normalidad de los residuos. En este caso, aunque se haya comprobado que los residuos del modelo son normales, no es posible aplicar los test de análisis de la homocedasticidad *Levenne* ni *Bartlett*. No es

apropiado aplicar el test de *Levenne* con variables cuantitativas. El test de *Bartlett* por otro lado, no se puede aplicar para cada modelo puesto que en el conjunto de datos *data.oxt* existe una observación para cada grupo de la variable que describe el tipo de estímulo (*stimulus.type*), cuando debería haber mínimo dos grupos por cada observación para poder aplicar el test correctamente.

2.3.5.3 Autocorrelación:

Para analizar la autocorrelación de los residuos del modelo, se ha utilizado el test de *Durbin-Watson*, que su hipótesis nula se define como la no autocorrelación (infiriendo independencia) entre los residuos y la hipótesis alternativa determina que sí existe correlación. Para aplicar este test, es necesario verificar que los residuos se distribuyen de forma normal, lo cual se ha comprobado anteriormente y por lo tanto sí que es posible aplicar el test mediante la función *durbinWatsonTest* sobre el modelo *mod.oxt2*.

Se observa que el p-valor es superior al 5% (p-valor=0.754) del nivel de significancia establecido, por lo tanto se asume que los residuos del modelo son independientes, ya que no hay evidencia suficiente para rechazar la hipótesis nula. Cabe recordar que en el diseño del modelo se ha eliminado la variable que mide el nivel de ritmo cardiaco post aplicación del estímulo (*hr.post*) puesto que estaba altamente correlacionada con el valor de ritmo cardiaco previo (*hr.bas*).

2.3.5.4 Multicolinealidad:

La multicolinealidad se obtiene cuando dos variables explicativas o más en un modelo de regresión múltiple están relacionadas linealmente. En este caso se analiza mediante el test de Farrar - Glauber si existe multicolinealidad entre las variables predictoras del *mod.oxt2*. Dado que todos los valores del *Klein* se igualan a cero, se asume que no se ha detectado colinearidad mediante el test de Farrar - Glauber. Otro método para calcular la multicolinealidad es utilizar la función *vif* del paquete *car*. La función *vif* - *Variance inflation factor* cuantifica la correlación entre las variables predictoras de un modelo, y se utiliza para analizar la colinearidad o la multicolinearidad entre las variables del modelo. Los valores más elevados significan que la correlación de esa variable con otra variable predictora del modelo será más alta, y normalmente valores superiores a 4 y 5 están considerados elevados, pero esto depende de cada caso. De las cuatro variables predictoras del modelo *mod.oxt2*, se obtienen valores cercanos a uno para todas ellas (mínimo 1.01 y máximo 1.25), por lo tanto cercanas a cero y por ello, suficiente para rechazar el principio de multicolinealidad en los residuos del modelo *mod.oxt2*.

2.3.6 Conclusión modelo Oxitocina

De los cuatro modelos que se han planteado para predecir el nivel de oxitocina tras aplicar un estímulo sobre los modelos (*mod.oxt2* explicado en la memoria y *mod.oxt*, *mod.oxt3*, y *mod.oxt4* descritos en el Anexo B), se ha demostrado que el modelo que mejores resultados ofrece es *mod.oxt2*, ya que, aunque no sea el único que cumple con todas las suposiciones para los residuos de un modelo lineal, sí que es el que obtiene un valor de R^2 ajustado más elevado. Además, es el único modelo donde todas las variables predictoras son significativas al 5%. Sin embargo, no es la única razón, ya que tras aplicar diferentes métodos de comparación de modelos (Anova, AIC o BIC), también es el con el que mejor ajuste se ha obtenido para los valores observados. Sin embargo, cabe destacar que el modelo *mod.oxt* ha quedado excluido de la comparación de modelos, puesto que no cumple con la suposición de homocedasticidad (tal y como se explica en el Anexo B con más detalle) para con los residuos de un modelo lineal. Por lo tanto, el modelo *mod.oxt2* se ha comparado con el modelo tercero y cuarto, utilizando Anova, AIC y BIC.

En la comparación Anova entre los modelos *mod.oxt2* y *mod.oxt3*, donde se busca obtener el valor RSS (*Residual Square Error* en inglés) más bajo, se observa que el valor de RSS es superior en el modelo *mod.oxt3* que en el *mod.oxt2*. Aplicando el método Akaike mediante las funciones AIC y BIC entre ambos modelos, donde se busca obtener el coeficiente más bajo en ambos casos (ya que demuestra un mejor ajuste del modelo), se ha obtenido un valor AIC = -13.94 y BIC=-2.97 para el modelo *mod.oxt2*, frente a un valor AIC = 6.82 y BIC=17.79 en el modelo *mod.oxt3*. Por lo tanto, aparte del valor de R^2 superior del modelo dos y de la significancia de la variable edad comentada previamente, existe evidencia suficiente para elegir el modelo *mod.oxt2* frente al modelo *mod.oxt3*. Para la comparación entre el modelo *mod.oxt2* y *mod.oxt4*, se aplica una vez más el método Akaike con las funciones AIC y BIC. En ambos casos, se obtiene valores más bajos para el modelo *mod.oxt2* que para el modelo *mod.oxt4* (AIC=42.7 y BIC=53.67), por lo que en este caso también se elige el segundo modelo frente al cuarto.

Finalmente, se concluye que con el número de observaciones incluidos en el estudio, el modelo más adecuado en predecir el nivel de oxitocina tras someter a una persona a un estímulo estresante es el modelo *mod.oxt2*. La ecuación, incluyendo los coeficientes de cada covariable es la siguiente:

$$\log(Y) = -1.325 - 0.607 \log(X_1) - 0.168 X_2 + \log(X_3) + 0.844 \log(X_4) + \epsilon$$

Ecuación 2: ecuación final incluyendo los coeficientes de cada covariable para describir el modelo mod.oxt2 y predecir el nivel de oxitocina tras aplicar un estímulo sobre el participante, transformando logarítmicamente las covariables numéricas y la variable respuesta

Siendo cada término,

- $\log(Y)$: variable respuesta *oxt.post* transformada logarítmicamente.
- -1.325: constante del modelo (B_0)
- $\log(X_1)$: variable predictora *age* transformada logarítmicamente.
- X_2 : variable categórica predictora *stimulus.type*.
- $\log(X_3)$: variable predictora *oxt.pre* transformada logarítmicamente.
- $\log(X_4)$: variable predictora *hr.bas* transformada logarítmicamente.

2.4 Biomarcador II: Cortisol

Para plantear el modelo que prediga el nivel de cortisol tras someter a una persona a un estímulo, lo primero que se ha hecho ha sido separar la base de datos principal y eliminar aquellas variables relacionadas con la oxitocina. Para ello se ha utilizado la función *select* del paquete *dplyr*. Las variables que se han eliminado han sido *-PANSS_general*, *-PANSS_negative*, *-PANSS_positive*, *-oxt.meas*, *-oxt.pre*, *-oxt.post*, *-arousal_level* y *-valence_level*. Finalmente, la base de datos generada para el análisis del cortisol se ha denominado *data.co* y está compuesta en un principio por 84 observaciones y 15 variables, que son las siguientes: *id*, *age*, *gender*, *disease*, *med.type*, *med.dos*, *oral.count*, *stimulus.type*, *co.meas*, *co.pre*, *co.post*, *co.reac*, *co.res*, *hr.bas* y *hr.post* (explicadas y descritas en la Tabla 3). Sin embargo, es necesario realizar un análisis de los datos para observar el comportamiento de las variables y ver si es necesario mantener todas ellas en el conjunto de datos a la hora de plantear el modelo.

2.4.1 Variable respuesta

La variable respuesta del modelo que se planteará en las siguientes secciones es *co.post*, que analiza el nivel de cortisol libre tras aplicar un estímulo sobre el participante. Esta variable se ha definido en la Tabla 3 y se trata de una variable cuantitativa continua.

Para obtener una descriptiva general de la variable, en la siguiente figura (Figura 13) se muestra un gráfico de cajas que describe su comportamiento:

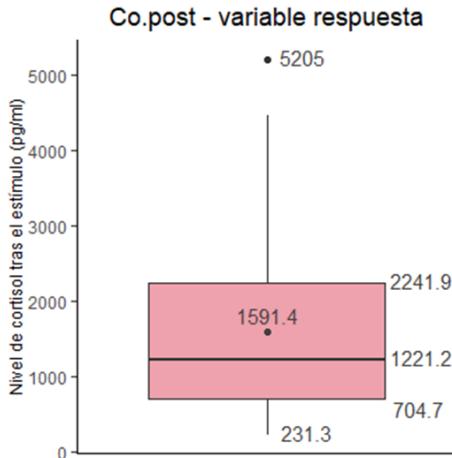


Figura 13: boxplot de la variable cortisol tras aplicar un estímulo sobre el participante, donde se muestran los valores de la media, mediana Q1, Q3, min y max, utilizando el conjunto de datos completo

En el gráfico se observa que la variable respuesta podría estar sesgada, y que tiene un valor *outlier* (influyente), que hace referencia al valor máximo de la variable en el conjunto de datos, con un valor de 5205.0 pg/ml, tal y como se observa en la siguiente Tabla 11. Además de este valor, en la tabla se recogen otros valores significativos de la variable que mide el nivel de cortisol tras aplicar un estímulo, *co.post* (el valor mínimo, la mediana, la media -junto con la desviación estándar- y los cuantiles Q1 y Q3). La media de los participantes es de 1591.4 pg/ml, con una desviación estándar de 1140.5.

Tabla 11: descriptiva numérica de la variable respuesta *co.post* (nivel de cortisol tras aplicar un estímulo sobre el participante) y valores de las medidas de dispersión

	Co.post
Valor general	
<i>Min</i>	231.3
<i>Q1</i>	704.7
<i>Mediana</i>	1221.2
<i>Media (SD)</i>	1591.4 (1140.5)
<i>Varianza</i>	1300770
<i>Q3</i>	2241.9
<i>Max</i>	5205.0
<i>Rango</i>	4973.7
<i>IQR</i>	1537.1

Tal y como se ha llevado a cabo para la variable de la oxitocina, mediante la función *describe* del paquete *dlookr*, se analiza la distribución de la variable respuesta del cortisol (*co.post*). Para la columna de *skewness*, la cual analiza la distribución simétrica de las observaciones, se obtiene un valor de 1.04, que es el mismo valor que se ha obtenido para la misma observación en la variable respuesta *oxt.post* del análisis anterior. En este caso, basándonos en el resultado numérico, no se considera que la variable se aleje demasiado del valor nulo, y por lo tanto parece que la variable está distribuida de manera normal, aunque esto se deberá analizar mediante diferentes tests que se llevarán a cabo posteriormente. Del gráfico en la Figura 13, se puede intuir que la variable está ligeramente sesgada a la derecha, debido a la distribución del tercer cuartil. El valor *outlier* observado en la figura anterior no parece que vaya a suponer un problema, puesto que para valor de *kurtosis* (que mide la influencia de los valores *outliers*) los valores cercanos a cero no suponen un problema, y en este caso se obtiene un valor de 0.47.

Para analizar si la variable sigue una distribución normal, se aplica el test de Shapiro-Wilk (con un nivel de significancia del 5%) tal y como se ha hecho para la variable de la oxitocina, donde la hipótesis nula del test acepta la distribución normal de los datos. En este caso, para la variable respuesta *co.post*, se obtiene un p-valor significativo ($6.19 \cdot 10^{-6}$), por lo tanto existe evidencia suficiente para no aceptar la hipótesis nula y considerar que la variable no sigue una distribución normal. El comportamiento de la variable se observa de forma gráfica en la siguiente imagen (Figura 14), donde se observa que para la variable *original* (es decir, sin llevar a cabo transformaciones sobre ella), claramente no se obtiene una distribución normal y además la variable está sesgada a la derecha. Además, el gráfico *Q-Q plot*, muestra que las diferentes observaciones de la variable no están sobreuestas en la línea continua diagonal, mostrando una vez más la falta de normalidad. De las dos transformaciones que se muestran (logarítmica y *sqrt*), es la primera la que más podría asemejarse a una distribución normal, aunque tampoco se podría afirmar únicamente observando el gráfico. Por lo tanto, se aplica el test de Shapiro-Wilk, pero esta vez sobre la variable respuesta *co.post* transformada logarítmicamente, donde en este caso se obtiene un valor de *p* igualado a 0.17, y por lo tanto no habría evidencia suficiente para rechazar la hipótesis nula y en este caso sí que se aceptaría la distribución normal de los datos.

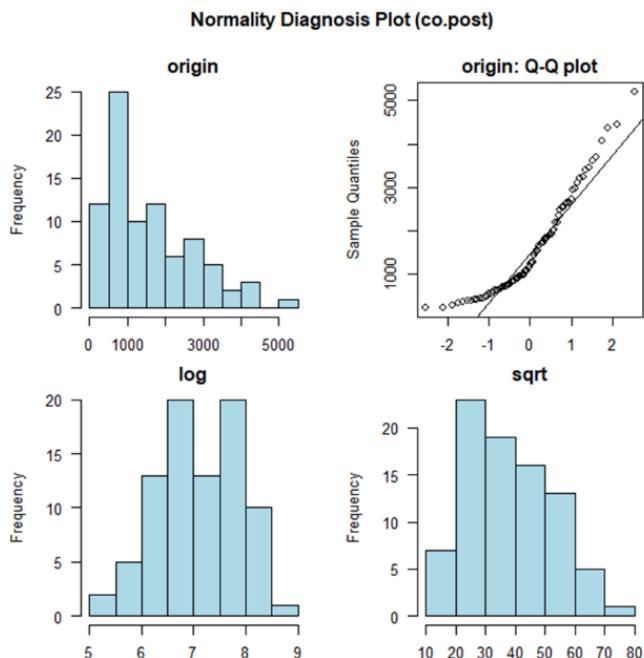


Figura 14: distribución de la variable respuesta que mide el nivel de cortisol tras aplicar un estímulo sobre el participante (*co.post*). Arriba a la izquierda, histograma de la distribución original. Arriba a la derecha, gráfico QQ de los datos originales. Los gráficos de abajo muestran histogramas de la distribución de la variable en caso de aplicar la transformación logarítmica o de raíz cuadrada a los datos. Análisis del conjunto de datos completo

2.4.2 Valores faltantes en el conjunto de datos

El conjunto de datos *data.co* está compuesto por 15 variables (incluyendo la variable respuesta (*co.post*) y 84 observaciones. Sin embargo, algunas variables presentan muchos valores faltantes (NA) en sus observaciones y esto podrá suponer un problema a la hora de plantear los modelos. Mediante la función *aggr* del paquete *VIM*, se visualiza en la Figura 15 la proporción de valores faltantes en el conjunto de datos (mostrados en la parte superior de la figura mediante barras), así como el gráfico las combinaciones para los valores faltantes (gráfico central).

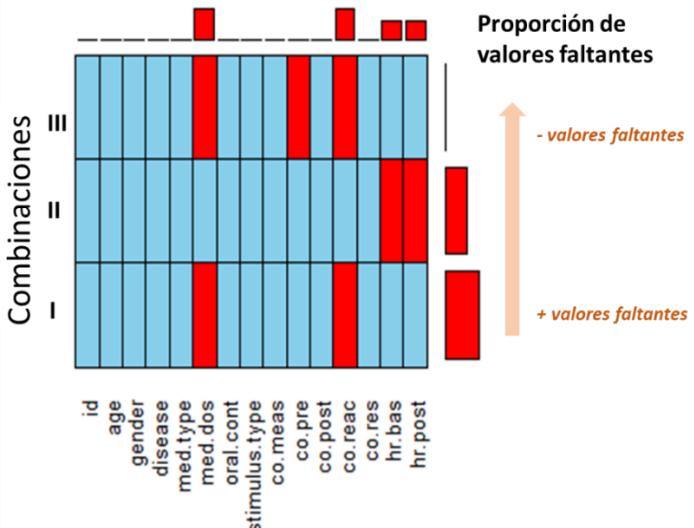


Figura 15: valores faltantes conjunto de datos cortisol para las variables numéricas, obtenido con la función aggr del paquete VIM. Proporción de valores faltantes (orden ascendente en la dirección marcada) para tres combinaciones

En la Figura 15, se muestra que una gran proporción de valores faltantes se encuentran en las variables *med.dos* (dosis de medicamento), *co.reac* (índice de reacción al cortisol, %) y *co.res* (respondedor o no al estímulo aplicado según el nivel de reacción). Sin embargo, para las dos primeras variables es posible imputar los *missings*: en el caso de la variable *med.dos*, para las observaciones donde los pacientes no toman medicación (*med.type* = 0), se puede imputar que la dosis será por lo tanto cero. La variable *co.reac* únicamente la calculan en el artículo de Tas et al. 2018 y la definen de la siguiente manera: cambio porcentual entre el nivel de cortisol previo y el cambio posterior al estímulo. Para ello, calculan la diferencia entre ambas mediciones de cortisol mediante las variables *co.pre* y *co.post* (*co.post* - *co.pre*), y posteriormente calculan el porcentaje de la diferencia respecto al nivel de cortisol previo. Por lo tanto, una vez conocida la función para calcular *co.reac*, es posible imputar estos valores también en las observaciones del estudio de Ooishi et al. 2017. Además, a partir de la variable *co.reac*, se pueden obtener los valores de *co.res* para las observaciones de Ooishi et al. 2017 donde esta variable se define como NA, ya que originalmente únicamente se calculan en el estudio de Tas et al. 2018, el cual se basa en el estudio de Miller et al. 2013 para clasificar a los pacientes como *responders* o *no responders*. La clasificación se define de la siguiente manera: aquellas observaciones con una reacción (*co.reac*) menor que el 15% relativa al nivel de cortisol previo no se considerarán *responders*, y los que tengan un porcentaje mayor, sí. Estos valores se han imputado en el conjunto de datos *data.co* utilizando funciones básicas del paquete *dplyr* como *mutate*, *select* o *filter*. Una vez imputados los *missings* en el conjunto de datos del cortisol, los valores faltantes se distribuyen de la siguiente manera, tal y como se muestra en la Figura 16:

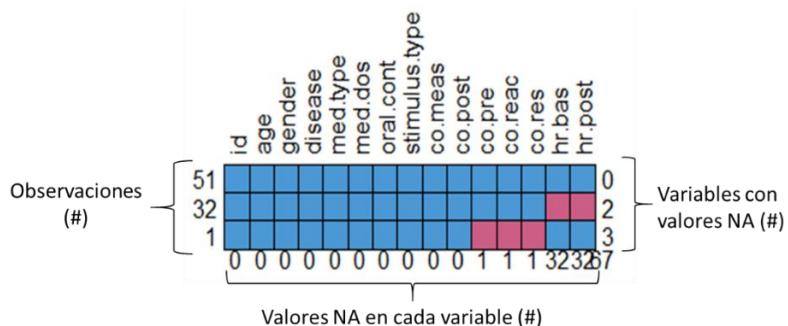


Figura 16: valores faltantes del conjunto de datos del cortisol *data.co*, donde hay 51 observaciones completas, 32 donde hay valores faltantes en dos variables y 1 observaciones con valores faltantes en tres variables. Hay 67 valores faltantes en total

De las 84 observaciones del conjunto de datos, 51 son observaciones completas, en 32 observaciones únicamente faltan las observaciones de las variables *hr.bas* y *hr.post*, y en una única observación falta la variable del cortisol previo al estímulo, y por lo tanto también faltan los valores en las variables *co.reac* y *co.res*. Las variables que mayor porcentaje de valores faltantes muestran son las que miden el ritmo cardiaco. Se ha consultado con diferentes expertos cuál debería ser el límite de valores faltantes aceptados para mantener una variable en el conjunto de datos, y el rango varía entre el 10 y el 30%. En este caso, para las variables que miden el ritmo cardiaco, el porcentaje de valores faltantes es del 38%. Sin embargo, de momento se decide mantener ambas variables, ya que añaden valor al estudio y en un futuro se pretende incluirlas en la recogida de muestras del proyecto en colaboración con la Universidad de Maryland. En las siguientes subsecciones (cuando se planteen los modelos y para el diseño de cada uno de ellos), se valorará si se deberán eliminar las 32 observaciones donde existen *missings* en las variables *hr.bas* y *hr.post* y por lo tanto trabajar sólo con casos completos. De momento, el conjunto de datos *data.co* tiene un total de 84 observaciones y 15 variables.

En este caso, a diferencia del análisis de la oxitocina, las variables categóricas *gender* (género), *disease* (existencia de enfermedad), *med.type* (tipo de medicamento), *stimulus.type* (tipo de estímulo) y *co.meas* (método en el que se ha medido el cortisol) tienen más de un nivel, por lo que todavía se mantienen en el conjunto de datos. Sin embargo, la variable *oral.count*, que mide la ingesta de anticonceptivos, debe eliminarse, puesto que tiene dos niveles: 0 o NA. Los valores NA para esta variable hacen referencia a los participantes masculinos, donde no tendría sentido preguntar si toman anticonceptivos orales, y los valores 0, se refiere a las mujeres participantes que no toman anticonceptivos orales. Dado que en ningún caso la variable está igualada a uno (ingesta del medicamento), esta variable se elimina del conjunto de datos. También se elimina del conjunto de datos la variable *id*, del mismo modo que se ha hecho para el análisis de la oxitocina. Por lo tanto, finalmente el conjunto de datos está compuesto por 84 observaciones y 13 variables.

2.4.3 Variables predictoras

De las 13 variables que componen el conjunto de datos, 12 serán variables predictoras en los modelos que se plantearán, ya que la variable restante es la dependiente. Algunas de las variables son numéricas (*age*, *med.dos*, *co.pre*, *co.reac*, *hr.bas* y *hr.post*) y el resto son categóricas (explicadas en la Tabla 3). Entre las variables categóricas, todas son factores de dos niveles, a excepción de la variable *stimulus.type*, que en este caso tiene tres (cada uno de ellos explicado en la Tabla 4 del documento). En la Figura 17, se muestran las variables categóricas *co.res* (respondedor o no), género y tipo de estímulo según cómo haya sido medida la muestra de cortisol. No se han incluido las variables categóricas de la enfermedad (*disease*) ni tampoco el del tipo de medicación, ya que como se ha comentado previamente, no hay variabilidad entre las variables en ambos tipos de muestras del biomarcador. Es decir, en las muestras de saliva, ninguna de las participantes está enferma ni toma medicación, y en la sangre todos los participantes padecen la misma enfermedad y toman la misma medicación. Las variables que se han incluido han sido porque sí que muestran variabilidad (o más de un nivel) en alguno de los dos grupos: saliva o sangre.

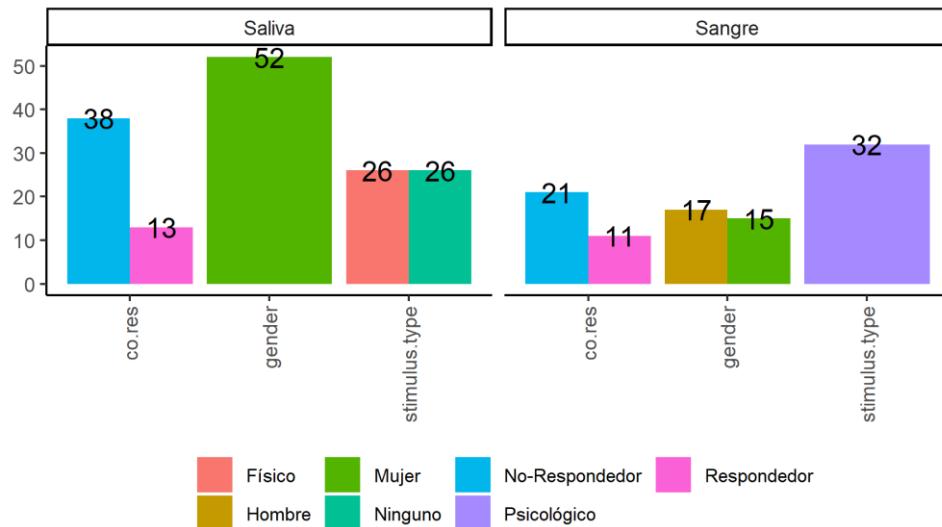


Figura 17: gráfico de barras de las variables categóricas co.res, gender y stimulus.type, que miden si el participante muestra un cambio o no en el nivel de cortisol tras el estímulo aplicado, el género del participante y el tipo de estímulo aplicado. Se utiliza el conjunto de datos del cortisol completo.

En la figura anterior se observa que para las muestras recogidas en la saliva, todas las participantes son mujeres ($N=52$), y que sin embargo, en el caso de las muestras de sangre, la muestra está nivelada según el género de los participantes. El tipo de estímulo, tal y como se ha ido comentando a lo largo del estudio, varía según el tipo de muestra que se ha cogido, por ello en el conjunto de datos de la saliva ambos tipos de estímulo (ninguno y físico) están igualados, y en la sangre, únicamente hay una barra, la cual se refiere al tipo de estímulo psicológico para generar estrés en los participantes. Sin embargo, la variable co.res, respondedor o no del cambio en el biomarcador cortisol según el estímulo, sí que varía en ambos conjunto de datos. En ambos casos, son más los participantes que pertenecen al grupo de no-respondedores. En las siguientes tablas (Tabla 12 y Tabla 13) se muestra un resumen de las frecuencias de cada variable y nivel, tal y como se ha observado en la Figura 17. Cabe destacar que en la tabla referente a la saliva existe un valor faltante, y por ello la suma de todas las frecuencias mostradas tiene un total de 51 observaciones en lugar de 52.

Tabla 12: tabla de frecuencias de las variables categóricas del conjunto de datos del biomarcador cortisol donde las muestras se han recogido en la saliva. Entre paréntesis el %. *Existe un valor NA para el tipo de estímulo "ninguno"

Cortisol medido en SALIVA (N=51*)		Respondedor? (Co.res)	
Tipo de estímulo (stimulus.type)	Ninguno Físico	No Respondedor	Respondedor
		16 (31%) 22 (42%)	9 (17%) 4 (8%)

Tabla 13: tabla de frecuencias de las variables categóricas del conjunto de datos del biomarcador cortisol donde las muestras se han recogido en la sangre. Entre paréntesis el %.

Cortisol medido en SANGRE (N=32)		Respondedor? (Co.res)	
Género	Mujer Hombre	No Respondedor	Respondedor
		12 (38%) 9 (28%)	5 (16%) 6 (19%)

Del mismo modo que en los apartados anteriores se ha mostrado la variable respuesta, a continuación se muestra la distribución de las variables numéricas continuas según el tipo de estímulo aplicado sobre ellas.

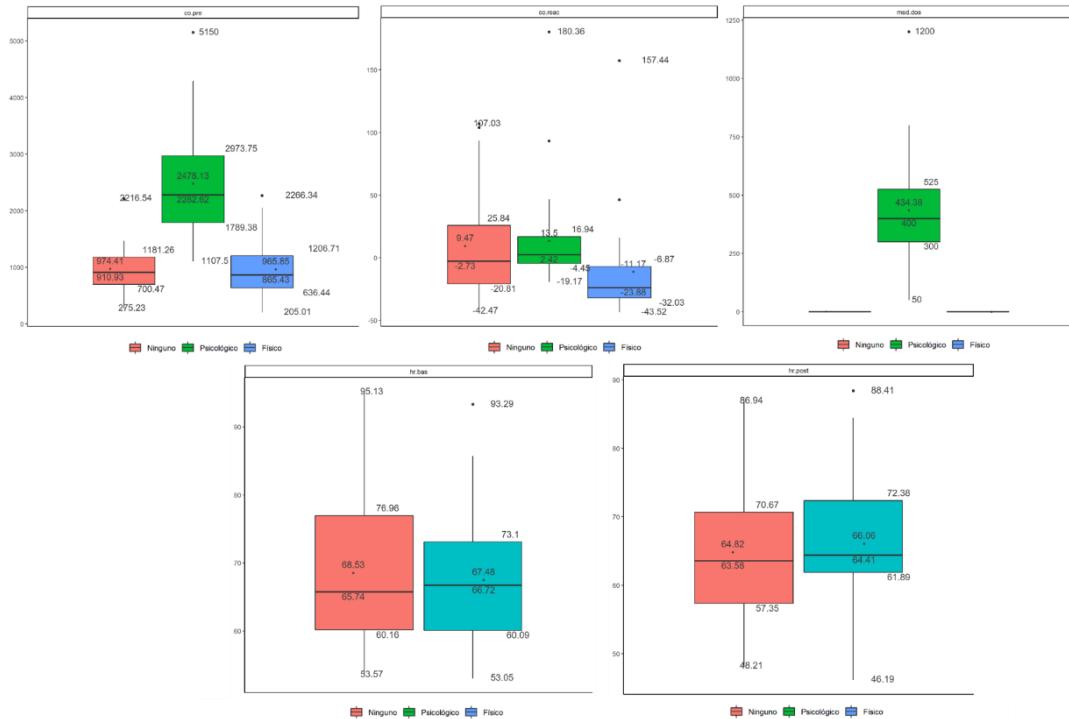


Figura 18: boxplots con los valores de la media, mediana, Q1, Q3, min y max para las variables numéricas del conjunto de datos del cortisol. Fila de arriba, izquierda a derecha: nivel de cortisol previo al estímulo, índice de reacción al cortisol y dosis del medicamento. Fila de abajo, izq. a dcha: ritmo cardiaco antes y después del estímulo aplicado en cada caso. Todos los gráficos están separados el tipo de estímulo que se aplique sobre el participante: ninguno, psicológico o físico

Como se ha mostrado en la Figura 16 del subapartado anterior respecto a los valores faltantes, no todas las variables tienen observaciones para cada tipo de estímulo. Es decir, como se muestra en la Figura 18 las variables que miden el ritmo cardiaco (*hr.bas* y *hr.post*) no se calculan para el tipo de estímulo psicológico, y por ello solo aparecen dos boxplots en la figura. De la misma manera, la dosis de medicación (*med.dos*) solo se mide para el tipo de estímulo psicológico y no para los otros dos, y por ello únicamente aparece un gráfico de cajas. Como se ha ido observando a lo largo del documento, esto depende del estudio original de donde se han cogido los datos para llevar a cabo el presente análisis. Las variables predictoras *co.pre* y *co.reac*, sí que se han medido para los tres tipos de estímulos (en el caso de la variable *co.reac* imputando los valores *missings* tal y como se ha explicado) y por ello aparecen los tres gráficos de cajas para ellos. En la siguiente Tabla 14 se recoge un resumen numérico de cada una de las variables, primero de forma general (variable general), y posteriormente separándola por los grupos (tipos de estímulos en este caso). La tabla se muestra a continuación:

Tabla 14: descriptiva numérica de las variables co.pre, co.reac, med.dos, hr.bas y hr.post, tanto de forma general como separandolas por el tipo de estímulo aplicado sobre ellas. Se recogen valores generales (min, max, media, mediana, Q1, Q3) y valores de las medidas de dispersión de cada una (varianza, rango, IQR)

	Variable				
	Co.pre	Co.reac	Med.dos	Hr.bas	Hr.post
Valor general					
Min	205.0	-43.52	00.00	53.05	46.19
Q1	797.5	-21.55	00.00	60.07	58.74
Median	1202.5	-2.73	00.00	66.01	63.71
Media (SD)	1551.48 (1023.12)	4.56 (41.25)	165.48 (272.76)	68.01 (10.31)	65.44 (9.47)
Varianza	1046775	1701.56	74398.02	106.30	89.68
Q3	2163.80	15.97	300.00	74.46	71.08
Max	5150.00	180.36	1200.00	95.13	88.41
Rango	4945	223.88	1200.00	42.08	42.22
IQR	1366.26	37.52	300.00	14.39	12.33
Ningún estímulo					
Min	275.23	-42.47	-	53.57	48.21
Q1	700.47	-20.81	-	60.16	57.35
Median	910.93	-2.73	-	65.74	63.57
Media (SD)	974.40 (478.07)	9.47 (43.28)	-	68.53 (10.74)	64.82 (9.73)
Varianza	228550.9	1873.16	-	115.35	94.67
Q3	1181.26	25.84	-	76.96	70.67
Max	2216.54	107.03	-	95.13	86.94
Rango	1941.31	149.5	-	41.56	38.73
IQR	480.79	46.65	-	16.80	13.32
Estímulo psicológico					
Min	1107.50	-19.17	50	-	-
Q1	1789.37	-4.45	300	-	-
Median	2282.62	2.42	400	-	-
Media (SD)	2478.13 (968.33)	13.50 (37.67)	434.37 (280.39)	-	-
Varianza	937663	1419.03	78618.55	-	-
Q3	2973.75	16.94	525	-	-
Max	5150.00	180.36	1200	-	-
Rango	4042.50	199.53	1150	-	-
IQR	1184.37	21.39	225.00	-	-
Estímulo físico					
Min	205.01	-43.52	-	53.05	46.19
Q1	636.44	-32.03	-	60.09	61.89
Median	865.43	-23.885	-	66.72	64.41
Media (SD)	965.85 (512.39)	-11.17 (40.49)	-	67.48 (10.04)	66.06 (9.35)
Varianza	262543.5	1639.44	-	100.80	87.42
Q3	1206.71	-6.87	-	73.10	72.38
Max	2266.34	157.44	-	95.29	88.41
Rango	2061.33	200.96	-	42.24	42.22
IQR	570.26	25.16	-	13.01	10.48

Otra variable predictora es la variable numérica *age*. Se trata de una variable discreta y la muestra utilizada para llevar a cabo este estudio, utiliza los mismos sujetos para cuando no se aplica ningún estímulo (*stimulus.type=0*) y cuando se aplica un estímulo físico (*stimulus.type =2*). Su distribución se muestra a continuación en la Figura 19, donde se puede observar que las cajas para dos de los estímulos son iguales.

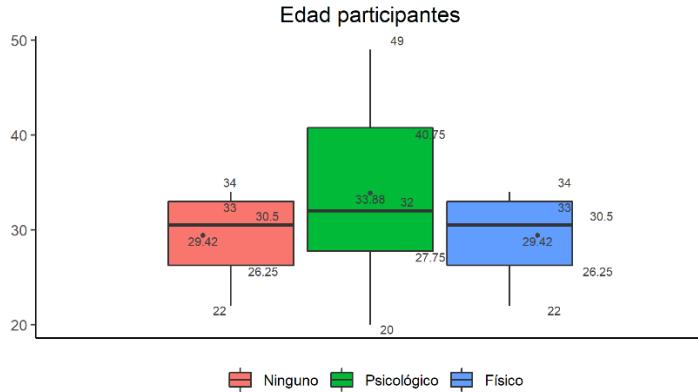


Figura 19: boxplot de la variable edad para cada tipo de estímulo del conjunto de datos *data.co* completo. Se muestran los datos numéricos del mínimo, máximo, media, mediana, Q1 y Q3 sobre el gráfico

De la misma manera que con las otras variables numéricas, en la Tabla 15 se muestra el resumen de los valores de la variable edad, tanto de forma general como separada por los tipos de estímulos.

Tabla 15: descriptiva numérica variable edad separada por el tipo de estímulo aplicado y de forma general, donde se recoge el valor mínimo, máximo, cuantiles, media, y valores de las medidas de dispersión (varianza, rango, IQR)

	Valor general	Estímulo =0, Estímulo =2	Estímulo =1
Edad			
Min	20.00	22.00	20.00
Q1	26.75	26.25	27.75
Median	31.00	30.50	32.00
Media (SD)	31.12 (6.37)	29.42 (4.11)	33.87 (8.30)
Varianza	40.58	16.89	68.89
Q3	34.00	33.00	40.75
Max	49.00	34.00	49.00
Rango	29.00	12.00	29.00
IQR	7.25	6.75	13.00

Para analizar el comportamiento general de las variables, es posible observar el valor de *skewness* para la simetría y el valor de *kurtosis* para los valores *outliers* de las variables numéricas como se ha hecho para el biomarcador oxitocina. En este caso, la variable cuyo valor de *skewness* es más alto es *co.pre* (nivel de cortisol previo al estímulo), con un valor de 2.08, el doble que el de la variable respuesta. Con el nivel de significancia establecido en un 5%, se analiza la normalidad mediante el test de Shapiro-Wilk de cada una de las variables, tal y como se ha llevado a cabo con la variable respuesta *co.post*, nivel de cortisol post estímulo.

Del test se obtiene que la variable que menos se asemeja a una distribución normal es la que mide la dosis del medicamento, *med.dos* (*p*-valor= $1.85 \cdot 10^{-12}$), aunque hay que tener en cuenta que muchas de las observaciones de esta variable eran originalmente valores NA, y que posteriormente se han transformado a valores nulos (igualados a cero), por lo tanto no es una variable que se espera vaya a tener un gran efecto en los análisis. La variable que le procede en

relación al p-valor para la distribución normal es *co.reac* (índice de reacción al cambio de cortisol), también con valores imputados para algunas de las observaciones. Finalmente, la variable que mide el nivel de cortisol previo al estímulo, *co.pre*, tiene un p-valor= $7.27 \cdot 10^{-6}$. Las únicas variables analizadas donde no existe evidencia suficiente para rechazar la hipótesis nula debido a que obtiene un p-valor superior al 5% es *hr.post*, que mide el ritmo cardiaco post estímulo. Es aconsejable analizar la distribución de las variables de forma gráfica para ver cómo se comportan y ver las posibles transformaciones para que se asemejen a la distribución normal, y para ello a continuación se muestran los gráficos obtenidos a partir de la función *plot_normality* para las variables *med.dos*, *co.reac*, *co.post*, *co.pre*, *age*, *hr.bas* y *hr.post*.

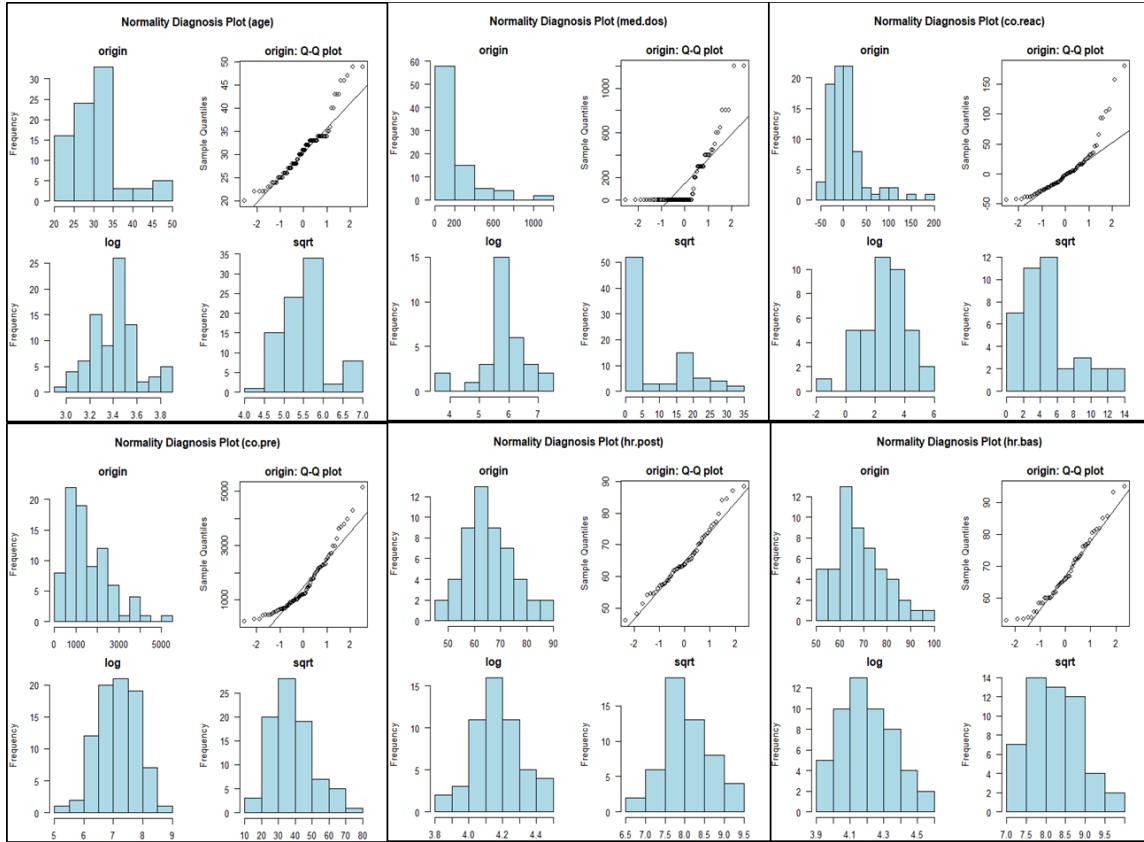


Figura 20: análisis de la normalidad. Fila arriba, izq. a dcha: variable edad, dosis ingerida, reacción del cortisol. Fila abajo, izq. a dcha: nivel de cortisol pre-estímulo, ritmo cardiaco post estímulo y ritmo cardiaco previo al estímulo
Para cada variable se muestra la distribución original mediante histograma y gráfico QQ, e histograma con transformación log y sqrt

Los outputs de la función *plot_normality* para cada una de las variables numéricas (Figura 20) confirman que el resultado que se observa gráficamente está relacionado con el p-valor analizado, ya que el histograma cuya distribución parece asemejarse a la normal sin aplicar ninguna transformación es únicamente el de la variable *hr.post* (aunque si la variable se transforma logarítmicamente, su p-valor aumenta de 0.27 a 0.85). Las variables del ritmo cardiaco previo (*hr.bas*) y cortisol previo (*co.pre*) están sesgadas a la derecha sin aplicar ninguna transformación, y sí que parece que al menos gráficamente su distribución mejora si son transformadas logarítmicamente. Si se analiza el p-valor de cada una con dicha transformación, se confirma que la distribución efectivamente mejora, obteniendo un p-valor=0.22 para *hr.bas* y p=0.70 para la variable *co.pre* y por lo tanto aceptando la hipótesis nula de normalidad según el test de Shapiro-Wilk. Se observa que para la variable *co.reac*, que mide el índice de reacción al cortisol, la mayoría de observaciones están comprendidas entre los valores de -50 y 50, y no parece que a simple vista la distribución de la variable se parezca más a una distribución normal al ser transformada. Sin embargo, el test de *normality* muestra un p-valor de 0.53 para la

transformación logarítmica de esta variable, por lo que sí se podría aceptar que se distribuya de forma normal tras ser transformada. Tal y como se ha comentado previamente, la variable que mide la dosis de medicamento (*med.dos*) es la que muestra un p-valor más bajo (debido en gran parte a la cantidad de valores nulos en las observaciones) y analíticamente al transformarla no se obtiene un p-valor superior al 5% (p-valor=0.01). Finalmente, la variable edad tiene más frecuencias en las primeras tres columnas, debido a que 26 pacientes son sometidos a dos de los tres tipos de estímulos en el estudio. A simple vista no parece que la variable edad siga una distribución normal en ninguno de los casos y analíticamente así lo demuestra la función *normality* con un p-valor=0.02 para su transformación logarítmica y 0.002 para la transformación de la raíz cuadrada, no aceptando por lo tanto la distribución normal con un nivel de significancia del 5% para esta variable predictora.

2.4.4 Análisis de la correlación de variables

Tal y como se ha llevado a cabo para el biomarcador I oxitocina, en este subapartado se realiza el análisis de la correlación para las variables que componen el conjunto de datos del cortisol. El objetivo es analizar si existen correlaciones lineales entre la variable respuesta y las variables predictoras, así como observar el comportamiento de las variables predictoras entre ellas. En este caso, a diferencia del análisis llevado a cabo para el biomarcador I, el conjunto de datos no está únicamente compuesto por observaciones completas, ya que se han mantenido algunos valores NA, y en algunas variables (referentes a los ritmos cardiacos sobre todo) el porcentaje de valores faltantes es elevado. Se ha aplicado sobre el conjunto de datos la función *cor*, con el método *Spearman*, puesto que se ha observado que no todas las variables cumplen con la normalidad antes de ser transformadas, y aplicando otro método (por ejemplo el de *Pearson*), el coeficiente de correlación podría variar si la variable fuera transformada posteriormente. Además, se ha igualado en el argumento “*use*” a “*pairwise.complete.obs*”, es decir, los valores faltantes se eliminan únicamente para realizar el cálculo de cada correlación por pares. Si se hubiera utilizado el argumento “*use*” igualado a “*complete.obs*”, la matriz de correlaciones estaría compuesta en su gran mayoría por valores NA, ya que con este argumento se eliminan todas las observaciones con algún valor faltante en ella. La matriz de correlaciones se muestra en la Tabla 16.

Tabla 16: matriz de correlación entre las variables que componen el conjunto de datos del cortisol (data.co) general, aplicando el método de Spearman

Coeficiente de correlación entre las variables conjunto de datos cortisol													
	age	gender	disease	med.type	med.dos	stimulus.type	co.meas	co.pre	co.post	co.reac	co.res	hr.bas	hr.post
age	1												
gender	-0.04	1											
disease	0.218	-0.642	1										
med.type	0.218	-0.642	1	1									
med.dos	0.234	-0.579	0.964	0.964	1								
stimulus	0	0	0	0	0	1							
co.meas	0.218	-0.642	1	1	0.964	0	1						
co.pre	0.264	-0.471	0.745	0.745	0.703	-0.025	0.745	1					
co.post	0.252	-0.464	0.774	0.774	0.724	-0.112	0.774	0.885	1				
co.reac	0.076	-0.167	0.331	0.331	0.284	-0.287	0.331	0.112	0.523	1			
co.res	0.039	-0.006	0.095	0.095	0.018	-0.18	0.095	0.051	0.392	0.785	1		
hr.bas	0.343	NA	NA	NA	NA	-0.038	NA	0.222	-0.012	-0.277	-0.304	1	
hr.post	0.419	NA	NA	NA	NA	0.097	NA	0.136	-0.057	-0.259	-0.287	0.862	1

Es deseable que la variable respuesta (*co.post*) esté relacionada con las variables predictoras que definirán el modelo. Por el contrario, no es deseable que las variables predictoras, las cuales deben ser independientes, estén correlacionadas entre ellas. En la Tabla 16 se observa la matriz de correlaciones, y para interpretar si la correlación es fuerte o débil, me he basado en los estudios de Martínez Ortega 2009 y Barrera 2014. El hecho de que el conjunto de datos del cortisol esté compuesto por los datos obtenidos de los estudios de Tas et al. 2018 y Ooishi et al. 2017 supone que los datos estén sesgados para analizar la correlación entre las variables que lo componen y esto queda en evidencia en los puntos que se describen a continuación.

- Las variables *disease* (enfermedad si o no), *med.type* (tipo de medicación), *med.dos* (dosis de medicación) y *co.meas* (método en el que se ha medido el cortisol) muestran una correlación perfecta entre ellas (*coef. = 1*). Con la variable *co.pre* (nivel de cortisol previo al estímulo) una correlación de 0.745 y con la variable respuesta *co.post* (*nivel de cortisol post estímulo*) un valor similar, 0.774, ya que la variable *med.dos* sólo se utiliza en el estudio de Tas et al. 2018 y para las observaciones del otro estudio, éstos valores se han imputado (igualándolos a cero, es decir, sin ninguna variabilidad). Obtener una correlación fuerte y positiva entre estas variables es debido una vez más al tipo de datos utilizados para el estudio. Todos los participantes que muestran una enfermedad (*disease=1*), toman medicación (*med.type=1*) y el nivel de cortisol ha sido medido en sangre (*co.meas=2*). Por el contrario, a los pacientes que no tienen una enfermedad y no toman medicación, la muestra se ha cogido en la saliva. Si la medición de cortisol hubiera estado aleatorizada entre esos pacientes (a algunos participantes muestra de saliva y a otros de sangre), el nivel de correlación entre las tres variables frente a *co.pre* y *co.post* sería más bajo y se hubiera evitado el patrón que se observa en el análisis.
- Ambas variables que miden el ritmo cardiaco (*hr.bas* y *hr.post*) muestran una correlación alta entre ellas, con un valor en el coeficiente de 0.862. Como en el caso del biomarcador oxitocina, se debe eliminar una de ellas a la hora de utilizarlas como variables predictoras en los modelos.
- Las variables *co.reac* (índice de reacción al cortisol) y *co.res* (respondedor o no según el índice) están correlacionadas de forma positiva y además con un valor muy alto (0.785). Es normal ya que *co.res* se genera a partir de los datos obtenidos en la variable *co.reac*.
- La variable *co.pre* y *co.post* (niveles de cortisol previo y posterior al estímulo aplicado) están altamente y positivamente correlacionadas entre ellas, con un coeficiente de correlación de 0.885 entre ambas variables. En la Figura 21 se muestra la correlación entre ambas.

Se ha comprobado que la matriz de correlación no difiere significativamente en el caso de que se hubiera utilizado el método de Pearson en el análisis, ya que las variables más correlacionadas seguirían siendo las mencionadas en los puntos anteriores.

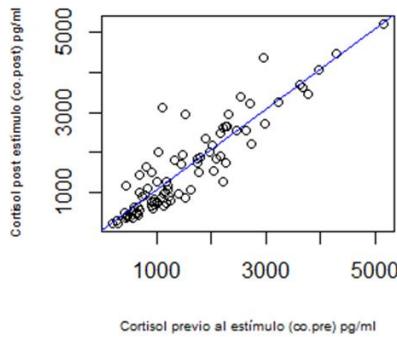


Figura 21: relación lineal entre la variable respuesta que mide el nivel de cortisol post estímulo (co.post) y la variable predictora del cortisol previo al estímulo (co.pre), donde se observa una gran correlación entre ambas variables

La correlación elevada entre variables predictoras supone que a la hora de plantear modelos, algunas de las variables que han mostrado una correlación alta con las demás covariables deban ser eliminadas, puesto que únicamente se deben incluir como variables predictoras las que muestren independencia entre ellas. Esto hará que los coeficientes con los que finalmente se plantee el modelo sean fiables. También es posible analizar la correlación entre las variables según el p-valor, y ver cuáles son significativos al 5%. En este caso se observa que se obtienen p-valores inferiores a 0.05 en las combinaciones que incluyen las variables *disease*, *med.type*, *med.dos* y *co.meas* (una vez más por el sesgo de los datos a raíz de los estudios utilizados), y también aquellas que incluyen la variable edad (ya que en uno de los estudios se aplican diferentes estímulos sobre un mismo paciente).

Finalmente, para concluir el análisis de la correlación, a continuación en la Figura 22 se muestra un mapa de calor (*heatmap*) donde se puede observar en color rojo las correlaciones más altas entre las variables. Tal y como se ha comentado en el presente subapartado, se observa que la interacción entre ambas medidas de ritmo cardiaco es alta, y que ocurre lo mismo en la interacción entre *co.res* y *co.reac* (tal y como se ha comentado previamente, *co.res* se genera a raíz de los valores obtenidos en *co.reac*) y también en la interacción *co.pre-co.post*, que miden los niveles de cortisol. Finalmente, la correlación más significativa se muestra entre las cuatro variables *co.meas*, *disease*, *med.dos* y *med.type*.

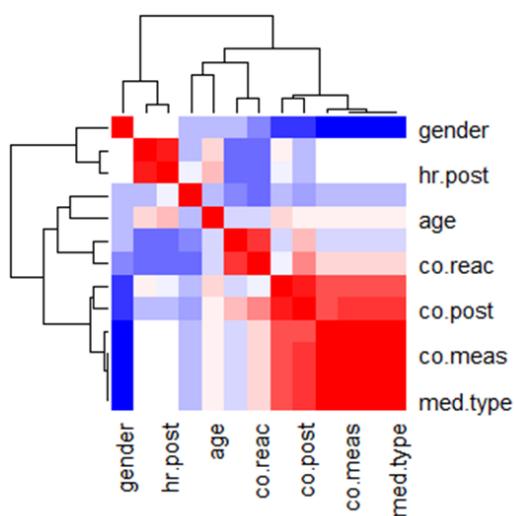


Figura 22: mapa de calor, heatmap para visualizar la correlación entre las variables del conjunto de datos *data.co* utilizando el conjunto de datos completo. Los rectángulos rojos identifican los coeficientes de correlación más cercanos a uno (más intensidad de rojo mayor correlación), y los rectángulos azules, menor correlación (mayor intensidad de azul menor correlación)

2.4.5 Modelo

Los coeficientes de correlación tan elevados obtenidos en el subapartado anterior limitan el diseño del modelo del cortisol. Como se ha comentado, las correlaciones tan altas se deben a que el conjunto de datos se ha generado a partir de la unión de dos bases de datos, donde cada una de ellas mide el cortisol de una forma diferente: mediante la saliva o mediante la sangre. Por ello, la variable *co.meas* (tipo de medición) está fuertemente relacionada con las variables *disease* y *med.type*, que claramente separan los datos según los estudios. Ocurre un fenómeno similar con la variable edad, ya que para el estudio donde las muestras se han medido en la saliva, a estos individuos se les han aplicado dos estímulos diferentes, y entonces cada uno de los participantes se repite en el conjunto de datos dos veces (es por ello por lo que los niveles de la variable *id* son 56 en lugar de 84), y eso hace que esta variable esté correlacionada con muchas de las variables que estaban en el conjunto de datos de ese estudio. Para poder trabajar con los datos pero a su vez asegurar la independencia entre las variables predictoras, se proponen dos posibilidades para plantear los modelos:

- 1) Con la variable respuesta *co.post*, limitar el modelo a aquellas variables del total del conjunto de datos que no estén correlacionadas. De este modo se obtendrá un modelo con el máximo de observaciones posible pero al mismo tiempo con menos variables predictoras que las analizadas para el conjunto de datos *data.co*.
- 2) Llevar a cabo un modelo por cada tipo de medición del cortisol. Se generará un modelo para las muestras obtenidas en la sangre y otro modelo para las muestras de saliva. Antes de llevar a cabo el modelo, en cada uno de los subapartados (saliva y sangre), se ha procesado un EDA del conjunto de datos final a utilizar ya que la distribución de algunas variables cambia al reducir el conjunto de datos.

2.4.5.1 Propuesta 1

Para la propuesta 1 se utiliza el conjunto de datos *data.co*, que está compuesto por 13 variables y 84 observaciones. A la hora de diseñar el modelo, se eliminan las variables que tienen un coeficiente de correlación más alto por pares, y sobre todo con la variable predictora *co.pre*, la cual indudablemente se incluye en el modelo ya que es la que mayor correlación tiene con la variable respuesta. Las variables que no se incluyen por lo tanto en el modelo son: *disease*, *med.type*, *med.dos*, *co.meas*, *co.res* y *hr.bas*. Entre las variables que miden el ritmo cardiaco, se ha elegido incluir la variable *hr.post*, ya que muestra un coeficiente de correlación más bajo frente a *co.pre* y la relación con la variable respuesta es similar entre ambas medidas del ritmo cardiaco. Sin embargo, el problema con las mediciones del ritmo cardiaco se da en los valores faltantes, ya que en el conjunto de datos hay 32 valores faltantes, y al pertenecer todas ellas a un estudio (y por lo tanto a un tipo de medición del cortisol), limita la variabilidad del modelo. Es por ello por lo que se decide eliminar la variable del modelo aunque su coeficiente de correlación con las otras variables no suponga un problema de independencia.

En el presente apartado, se analiza el modelo que mejores resultados ha mostrado para la predicción del cortisol, aunque no se haya cumplido con la suposición de la normalidad. Los tres modelos que también se han planteado y analizado en un principio se muestran en el Anexo C, aunque finalmente se hayan descartado, debido que se han obtenido peores resultados en los residuos y los valores de AIC y BIC han sido mucho más elevados que para el modelo *mod.co.p2* analizado en la presente subsección. De los tres modelos descartados, ninguno ha cumplido con el supuesto de homocedasticidad, puesto que se han obtenido p-valores para los test aplicados inferiores a 0.05. Sin embargo, en el primer modelo descartado, los residuos muestran mayor

linealidad que el modelo elegido, pero incluyen una covariable no significativa pese a haber aplicado *stepwise regression*, y además el supuesto de autocorrelación está en el límite, ya que tiene un p-valor igualado a 0.05.

Con todo ello, en la siguiente ecuación se presenta el planteamiento inicial del modelo que "mejores" resultados ha mostrado para la predicción del cortisol, donde se ha llevado a cabo una doble transformación logarítmica en las covariables numéricas y también en la variable dependiente.

$$\log(Y) = B_0 + B_1 \log(X_{age}) + B_2 \log(X_{co.pre}) + B_3 \log(X_{co.reac}) + B_4 X_{gender} \\ + B_5 X_{stimulus.type} + \epsilon$$

Ecuación 3: planteamiento inicial modelo con mejores resultados en la propuesta 1 para predecir el nivel de cortisol tras la aplicación de un estímulo sobre el participante. Conjunto de datos completo del cortisol, data.co. Variable dependiente y covariables transformadas logarítmicamente

En un principio, el modelo que se ha planteado tiene como variables predictoras *age*, *co.pre*, *co.reac*, *gender* y *stimulus.type*, transformando logarítmicamente las numéricas (tanto continuas como discretas). La variable respuesta, también se plantea con la misma transformación que las covariables. Con la doble transformación logarítmica, se ha obtenido una varianza más constante en los residuos del modelo y mayor normalidad en los residuos que en el caso donde no se ha aplicado ninguna transformación. En la Tabla 17 se muestra el *output* obtenido del sumario del modelo final tras reducir el número de variables en el modelo y dejando únicamente las significativas al 5%:

Tabla 17: resultados del modelo de regresión para predecir el nivel de cortisol tras aplicar el estímulo en el participante, con el logaritmo de las covariables co.pre y co.reac como predictores del nivel de cortisol

	Coeficiente B	Std.Err	t	Sig
Predictores				
Constante	-0.28715	0.27410	-1.048	0.302
log(co.pre)	1.01465	0.03431	29.576	< 2e-16 ***
log(co.reac)	0.15950	0.01782	8.948	1.11e-10 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
F	447.3			
R ²	0.9592			
p-valor	< 2.2e-16			

En la tabla anterior se puede observar que finalmente, las únicas variables que han resultado significativas al 5% han sido *log(co.pre)* y *log(co.reac)*. Aunque en un principio el modelo se haya planteado con las variables predictoras descritas anteriormente no todas han resultado significativas, y tras aplicar Akaike mediante la función *StepAIC*, se ha determinado que únicamente debían incluirse las dos variables mencionadas. El valor del *R²* es 0.9592, considerándolo un valor muy alto. Tras el planteamiento, es necesario analizar el comportamiento de los residuos del modelo, ya que en base a los resultados que muestren, se podrá determinar si los coeficientes obtenidos para cada variable son fiables o no, y por lo tanto, valorar si es posible estimar la variable respuesta con el presente planteamiento. A continuación, en la Figura 23, se muestran cuatro gráficos diferentes que describen los residuos del modelo *mod.co.p1*.

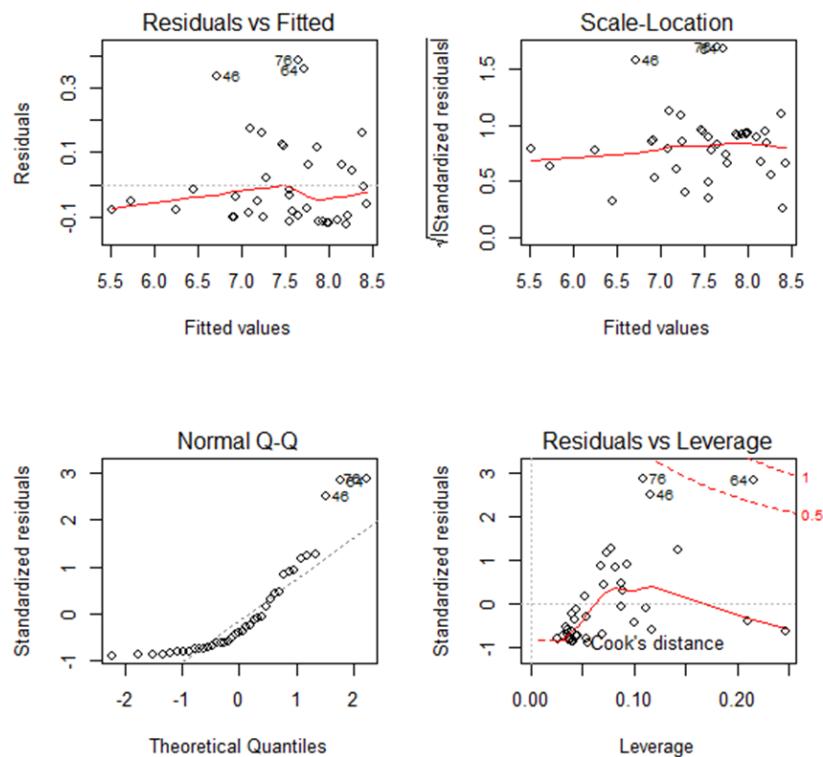


Figura 23: : distribución de los residuos del modelo *mod.co.p1* (transformación doble log sobre todas las covariables y la variable dependiente) para predecir el nivel de cortisol tras aplicar un estímulo. Gráfico de linealidad (arriba izq.), homocedasticidad (arriba dcha.), normalidad (abajo izq.) y puntos outliers o influyentes (abajo dcha.).

En la Figura 23 se observa mediante el gráfico *Scale Location* que parece que el modelo sí cumple la suposición de homocedasticidad, y que por lo tanto la varianza de los residuos está distribuida de forma constante, ya que la línea roja del gráfico es casi horizontal. Sin embargo, en los demás gráficos parece que la influencia de valores *outliers* es muy alta para los resultados de linealidad y normalidad. Se ha comprobado que de eliminar los valores *outliers*, sí que se conseguiría un valor más alto respecto al R^2 , pero que no mejoraría las suposiciones de linealidad ni normalidad gráficamente ni en los test aplicados. Por lo tanto, no se considera que eliminar los valores influyentes (en concreto las observaciones 46, 64 y 76 que se observan en el gráfico previo) del conjunto de datos sea efectivo en este caso.

Al aplicar el test de Shapiro-Wilk en los residuos del modelo, donde se quiere verificar si éstos siguen o no una distribución normal, se observa que el p-valor obtenido tiene un valor de $1.26 \cdot 10^{-5}$, por lo tanto se rechaza la hipótesis nula del test y no se asume la normalidad de los residuos. El no cumplir con la suposición de la normalidad ni de la linealidad (mostrada en la Figura 23), es suficiente para rechazar este modelo para predecir el valor del *co.post* con el conjunto de datos general del cortisol. Tal y como se ha mencionado previamente, ninguna de las transformaciones de los datos que se ha llevado a cabo (mostradas en el Anexo C) ha cumplido con la hipótesis de la normalidad y han mostrado peores resultados que el modelo analizado. Por lo tanto, se rechaza la propuesta 1 como posibilidad de predecir el nivel de cortisol utilizando un conjunto de datos con más observaciones, y se procede a la propuesta número 2, donde el cortisol se analiza dependiendo del método en el que se han recogido las muestras, pudiendo ser en sangre o en saliva en este estudio, tal y como se explica en los siguientes subapartados.

2.4.5.2 Propuesta 2

Para realizar los modelos según la propuesta número 2, la base de datos *data.co* se debe dividir en dos según el modo en el que se ha medido el biomarcador cortisol: en la saliva o en la sangre. Antes de plantear el modelo en cada uno de los subapartados de la sangre y la saliva, se lleva a cabo un EDA para conocer qué variables predictoras se deben incluir en cada conjunto de datos, la distribución de cada una de las variables, y también la correlación por pares entre las variables para el nuevo conjunto de datos en cada caso.

2.4.5.2.1 Sangre

Con el objetivo de generar el modelo utilizando únicamente observaciones de la sangre, se ha separado el conjunto de datos original *data.co* según los valores en la variable que mide el método de medición (*co.meas*). Este nuevo conjunto de datos se ha denominado *data.co.sngr*, y en un principio estará compuesto por 7 variables y 32 observaciones. En comparación con la base de datos original (*data.co*), se han eliminado seis variables: *disease* (ya que todos tienen la misma enfermedad), *med.type* (ya que todos toman la misma medicación), *stimulus.type* (a todos se les aplica el mismo estímulo), *co.meas* (todos se han medido en la sangre), y las variables *hr.bas* y *hr.post*, puesto que en el estudio de donde se han obtenido las observaciones en sangre no se ha medido el ritmo cardíaco de sus participantes. No existe ningún valor faltante en el conjunto de datos *data.co.sngr*. Aunque la distribución de los datos no variará mucho del análisis exploratorio llevado a cabo en los subapartados anteriores, dado que el número de observaciones ha disminuido, a continuación se vuelven a mostrar estas variables tanto gráficamente (Figura 24) como numéricamente en la Tabla 18. Finalmente, también se volverá a analizar la correlación entre variables, ya que en este caso, la reducción de la base de datos sí que podrá modificar los coeficientes de correlación entre las variables que componen el conjunto de datos.

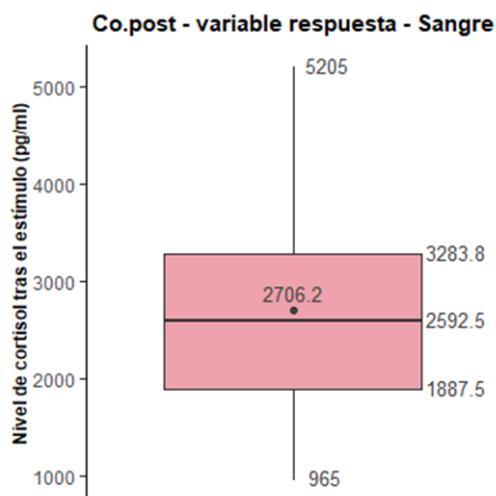


Figura 24: boxplot de la variable cortisol tras aplicar un estímulo sobre el participante, donde se muestran los valores de la media, mediana Q1, Q3, min y max, utilizando el conjunto de datos con las mediciones en sangre

Tabla 18: descriptiva numérica de la variable respuesta co.post (nivel de cortisol tras aplicar un estímulo sobre el participante) para el conjunto de datos con mediciones en sangre

	Co.post
Valor general	
<i>Min</i>	965
<i>Q1</i>	1887.5
<i>Mediana</i>	2592.5
<i>Media (SD)</i>	2706.25 (992.23)
<i>Varianza</i>	984520.4
<i>Q3</i>	3283.75
<i>Max</i>	5205
<i>Rango</i>	4240.00
<i>IQR</i>	1396.25

Para conocer la distribución de la variable respuesta *co.post* en el conjunto de datos, se vuelve a aplicar el test de Shapiro-Wilk mediante la función *normality*. Se obtiene un p-valor de 0.62, por lo tanto no existe evidencia suficiente para rechazar la hipótesis nula del test y se acepta la normalidad en la distribución de los datos de la variable respuesta. En la Figura 25 se vuelve a mostrar de forma gráfica el comportamiento de los datos, y a simple vista no parece que la transformación de los datos suponga una mejora en cuanto a la normalidad de se refiere en comparación con el original. Además, los puntos del gráfico Q-Q parecen que en general están sobreestimados en la línea de la normal, aunque en la cola haya unos puntos que difieren.

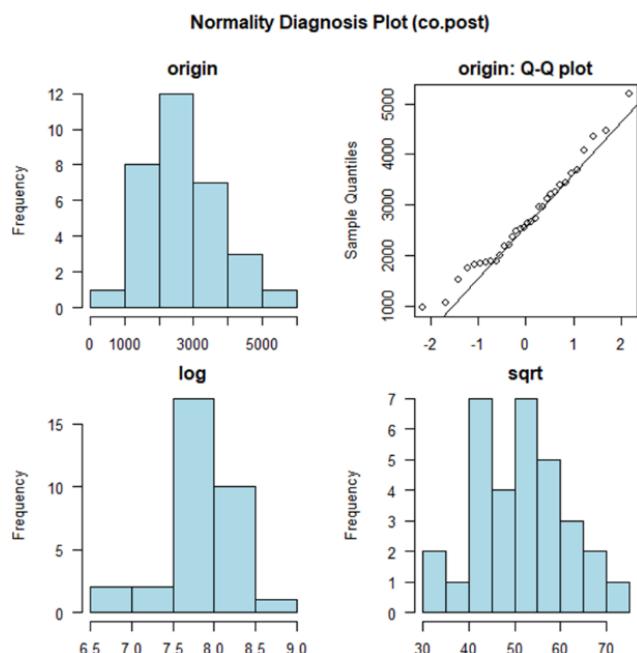


Figura 25: distribución de la variable respuesta que mide el nivel de cortisol tras aplicar un estímulo sobre el participante (*co.post*). Arriba a la izquierda, histograma de la distribución original. Arriba a la derecha, gráfico QQ de los datos originales. Los gráficos de abajo muestran histogramas de la distribución de la variable en caso de aplicar la transformación logarítmica o de raíz cuadrada a los datos. Conjunto de datos con las mediciones en sangre

Respecto a las variables predictoras, en la siguiente Figura 26 se muestra la distribución de las mismas, y en la Tabla 19 se resumen los datos más significativos de cada una de las variables para este conjunto de datos, aunque estos datos ya se han mostrado por grupos en las Tabla 14 y Tabla 15.

Tabla 19: descriptiva numérica de las covariables co.pre, co.reac, med.dos y age de forma general (estímulo psicológico). Se recogen valores generales (min, max, media, mediana, Q1, Q3) y valores de las medidas de dispersión de cada una (varianza, rango, IQR). Conjunto de datos del cortisol con mediciones en sangre

	Variable			
	Co.pre	Co.reac	Med.dos	Age
Valor general				
Min	1107.50	-19.17	50	20.0
Q1	1789.37	-4.45	300	27.75
Median	2282.62	2.42	400	32.0
Media (SD)	2478.13 (968.33)	13.50 (37.67)	434.37 (280.39)	33.87 (8.30)
Varianza	937663.0	1419.03	78618.55	68.89
Q3	2973.75	16.94	525	40.75
Max	5150.00	180.36	1200.0	49.0
Rango	4042.5	199.53	1150.0	29.0
IQR	1184.37	21.39	225.0	13.0

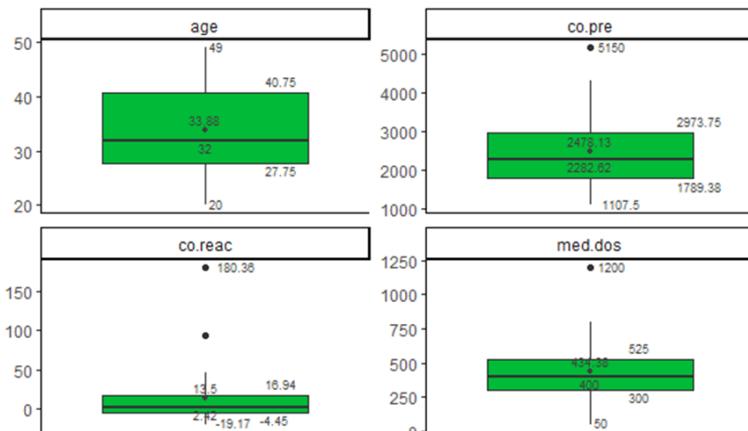


Figura 26: boxplots con los valores de la media, mediana, Q1, Q3, min y max para las variables numéricas del conjunto de datos con las mediciones de la sangre. Fila de arriba: variable edad y nivel de cortisol previo a la aplicación del estímulo. Fila abajo: nivel de reacción frente a los estímulos y dosis ingerida de los participantes. Todas las observaciones pertenecen al tipo de estímulo psicológico.

Respecto a la distribución normal de estas variables, sin aplicar ninguna transformación sobre ellas, la única variable significativa al 5% es *co.pre* que mide el nivel de cortisol previo al estímulo, con un p-valor ligeramente superior al 5% (p-valor=0.083) y por lo tanto se aceptaría la distribución normal para la variable. En la figura anterior se ha observado que no parece que esta variable esté sesgada, ya que la distribución en el gráfico de cajas parece muy similar tanto encima como debajo de la mediana. Si las variables se transforman logarítmicamente, la única variable no significativa al 5% es *med.dos* (dosis del medicamento), con un p-valor=0.01. La distribución de estas variables se muestra a continuación en la Figura 27:

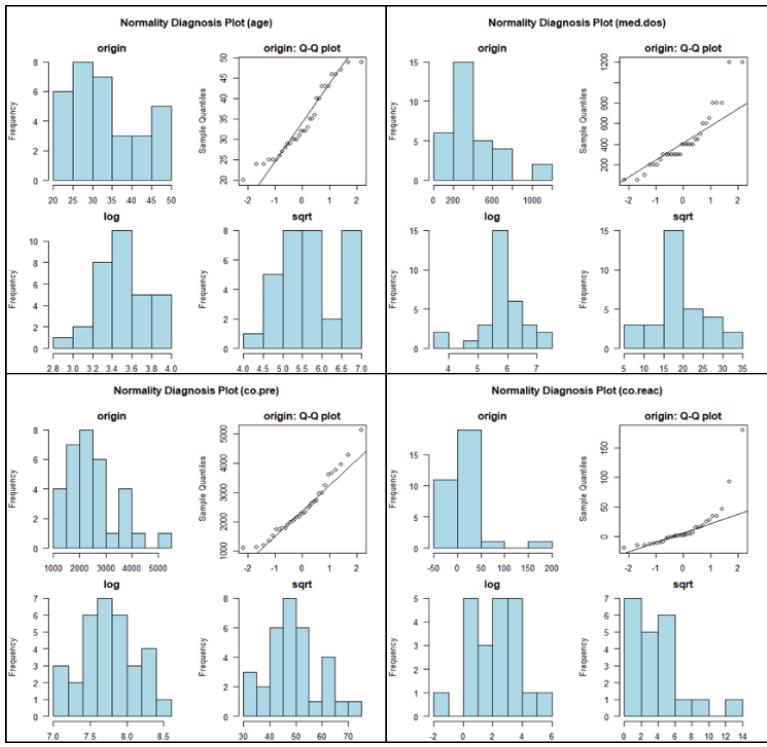


Figura 27: análisis de la normalidad. Fila arriba, izq. a dcha: variable edad y dosis ingerida. Fila abajo, izq. a dcha: nivel de cortisol previo al estímulo y reacción del cortisol frente a los estímulos. Conjunto de datos del cortisol medido en sangre. Para cada variable se muestra la distribución original mediante histograma y gráfico QQ, e histograma con transformación log y sqrt

Finalmente, respecto a las correlaciones entre las variables, en la Tabla 20 se muestran los valores de los coeficientes de correlación para los datos del conjunto de datos de la sangre. Se observa que los coeficientes de correlación más altos se dan entre las variables *co.res* y *co.reac* y también entre *co.post* y *co.pre*, una tendencia que ya se ha ido observando en los análisis de correlaciones previos. Estos resultados se reflejan también en el mapa de calor de la Figura 28.

Tabla 20: matriz de correlación para las variables del conjunto de datos del cortisol medido en sangre

Coeficiente de correlación para las variables del conjunto de datos de la sangre

	age	gender	med.dos	co.pre	co.post	co.reac	co.res
age	1.00						
gender	0.19	1.00					
med.dos	0.15	0.20	1.00				
co.pre	0.18	0.05	-0.09	1.00			
co.post	0.16	0.17	-0.18	0.80	1.00		
co.reac	0.14	0.15	-0.31	-0.14	0.34	1.00	
co.res	0.14	0.11	-0.42	-0.22	0.19	0.82	1.00

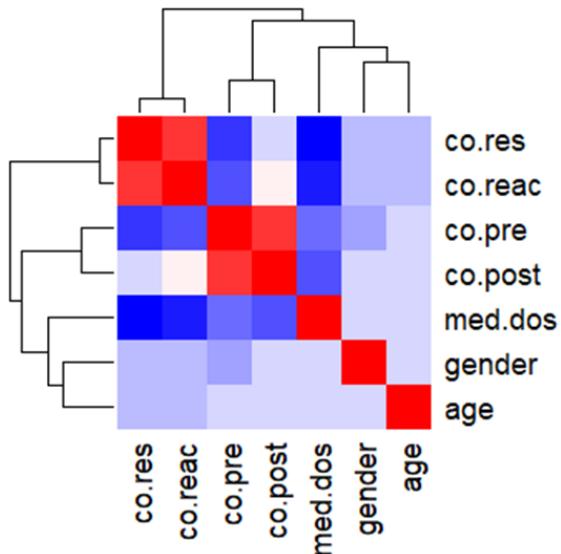


Figura 28: mapa de calor (heatmap) a partir de los coeficientes de correlación para las variables del conjunto de datos del cortisol medida en sangre. Los rectángulos rojos identifican los coeficientes de correlación más cercanos a uno (más intensidad de rojo mayor correlación), y los rectángulos azules, menor correlación (mayor intensidad de azul menor correlación)

Una vez conocidos los datos de este conjunto de datos, se procede a explicar el modelo con el que se han obtenido mejores resultados.

Modelo sangre - cortisol

En el subapartado donde se ha analizado la correlación se ha observado que las variables con mayor correlación por pares son *co.res* (respondedor o no al cortisol) y *co.reac* (índice de reacción para el cortisol) para el conjunto de datos de la sangre. Para el diseño de los modelos, se ha mantenido la variable *co.reac* en lugar de *co.res*, por tratarse de una variable numérica y no una variable categórica, y porque la variable *co.res* se genera en función de los valores en la variable *co.reac* (para los valores superiores al 15% en la variable *co.reac*, *co.res* se iguala a dos, y si no, se iguala a uno).

Para el desarrollo de este apartado, se han planteado cuatro modelos diferentes. En la presente sección se muestra el modelo con mejores resultados para predecir la variable respuesta y en el Anexo D del presente documento los otros modelos planteados pero finalmente descartados, ya que ninguno de ellos ha cumplido el supuesto de normalidad y homocedasticidad para los residuos. El modelo que se plantea a continuación se denomina *mod.co.sngr3*, la fórmula que se ha planteado es la siguiente y tal y como se observará más adelante, éste modelo sí que cumple ambos supuestos (además de otros) para sus residuos:

$$\log(Y) = B_0 + B_1 X_{co.pre} + B_2 X_{age} + B_3 X_{co.reac} + B_4 X_{med.dos} + \epsilon$$

Ecuación 4: planteamiento inicial del modelo *mod.co.sngr3* utilizando el conjunto de datos del cortisol para las mediciones en sangre y predecir el nivel de cortisol tras la aplicación del estímulo. Variable dependiente *co.post*, transformada logarítmicamente.

El modelo está compuesto por las variables predictoras *co.pre* (nivel de cortisol previo al estímulo), *age* (edad), *co.reac* (índice de reacción al cortisol) y *med.dos* (dosis de medicamento), y la variable respuesta (*co.post*) transformada logarítmicamente, ya que de este modo ha resultado cumplir las hipótesis del modelo de regresión (en concreto respecto a normalidad y homocedasticidad de los residuos) y las demás transformaciones no lo han hecho. En un primer planteamiento, se había incluido la variable predictora *gender*, pero tras aplicar la función *stepAIC* para llevar a cabo la selección de los predictores del modelo, se ha eliminado, ya que no

era significativa y por lo tanto no tenía un efecto sobre la variable respuesta *co.post*. En la Tabla 21 se muestra el *output* obtenido del modelo:

Tabla 21: resultados del modelo de regresión, variables *co.pre*, *age*, *co.reac* y *med.dos* como predictores del nivel de cortisol post la aplicación de la situación de estrés

Predictores	Coeficiente B	Std.Err	t	Sig
Constante	6.745e+00	8.432e-02	79.985	< 2e-16 ***
<i>co.pre</i>	3.852e-04	1.843e-05	20.904	< 2e-16 ***
<i>age</i>	4.926e-03	2.071e-03	2.379	0.024691 *
<i>co.reac</i>	5.388e-03	4.816e-04	11.189	1.21e-11 ***
<i>med.dos</i>	-2.408e-04	6.244e-05	-3.857	0.000645 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
F	125.5			
R ²	0.9414			
p-valor	< 2.2e-16			

En la Tabla 21 se observa que el valor de R^2 ajustado es 0.9414, y que todas las variables predictoras son significativas al 5%. Tras el planteamiento, es necesario analizar el comportamiento de los residuos del modelo, ya que en base a esos resultados, se podrá determinar si los coeficientes obtenidos para cada variable son fiables o no para estimar el valor de la variable respuesta. A continuación, en la Figura 29, se muestran cuatro gráficos que describen los residuos del modelo *mod.co.sngr3*.

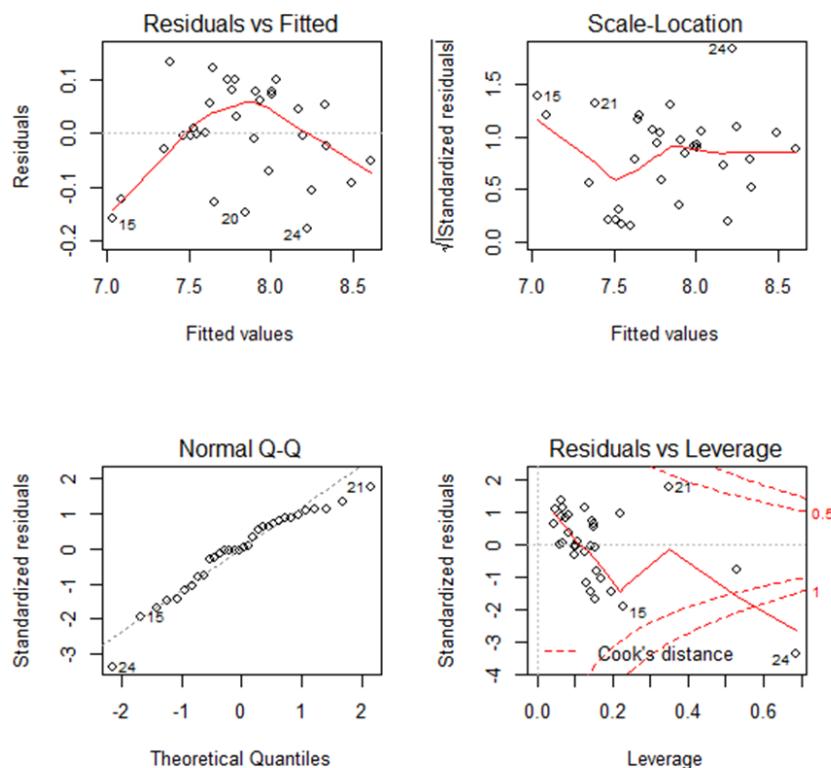


Figura 29: distribución de los residuos del modelo *mod.co.sngr3* (transformación logarítmica de la variable respuesta). Gráfico de linealidad (arriba izq.), homocedasticidad (arriba dcha.), normalidad (abajo izq.) y puntos outliers o influyentes (abajo dcha.)

Cada uno de los gráficos mostrados analiza diferentes aspectos de los residuos del modelo, descritos a continuación:

- Linealidad: analizado en el gráfico *Residuals vs Fitted*, que muestra si el modelo es una combinación lineal de las variables predictoras. En este caso, no parece que los residuos se distribuyan alrededor de la línea horizontal de manera homogénea, puesto que la línea roja que marca la distancia mínima entre los residuos no es horizontal y no se distribuye encima de la línea marcada en el valor cero. Aunque la linealidad a simple vista no parece que se cumpla, se sigue analizando el modelo para las otras suposiciones.
- Normalidad: analizado en el gráfico *Normal Q-Q*, que muestra si los residuos están distribuidos de forma normal. Para que se considere que los residuos están distribuidos de forma normal, éstos deberían estar encima de la línea discontinua. En este caso, se observa que en las colas hay algunos valores que difieren de la línea, lo que sugiere que pueden haber valores *outliers*. Sin embargo, la mayoría de observaciones sí que está encima de la línea discontinua central, por lo que a simple vista sí que se podría aceptar la hipótesis de normalidad de los residuos.
- Homocedasticidad: analizado en el gráfico *Scale Location*, que muestra si la varianza de los residuos está distribuida de forma constante para las variables predictoras. En este caso se observa que la línea roja no es horizontal pero tampoco tiene una forma acampanada, por lo que hay poca evidencia gráfica para ver si los residuos son homocedásticos, o por el contrario heterocedásticos. Se aplicarán diferentes tests para analizar este supuesto.
- Detectar valores influyentes (*outliers*) del modelo: mediante el gráfico *Residuals vs Leverage*. Los valores que se muestran separados del resto mediante la línea discontinua, son valores influyentes, que de eliminarlos, el comportamiento del modelo cambiaría, aunque se ha analizado que no mejoraría. Se ha llevado a cabo un análisis eliminando las observaciones número 15, 21 y 24 (que son las más distanciadas del resto y también más cercanas a distancias más altas de Cook). El modelo seguiría cumpliendo las mismas características que las analizadas mediante los diferentes tests, pero se volverían a generar nuevos valores influyentes en los residuos del modelo, algunos incluso más distanciados que los que se han observado, indicando una vez más que se podrían tratar como valores distanciados más que valores influyentes en el modelo. No se considera que de eliminarlos el modelo mejore, ya que gráficamente el comportamiento de la linealidad de los residuos es similar con y sin los puntos mencionados y también la normalidad empeora al haber eliminado observaciones del conjunto de datos. Finalmente, la variable edad dejaría de ser significativa al 5%, y de eliminarla como covariable, los residuos tendrían peores comportamientos. Por lo tanto, no se considera que eliminar los valores mencionados sea necesario para el desarrollo del presente modelo.

Para corroborar los supuestos analizados gráficamente, tal y como se ha comentado se aplican diferentes tests, mostrados en los siguientes subapartados.

- **Normalidad de los residuos:**

Lo primero que se deberá hacer será verificar mediante un test de normalidad si los residuos del modelo *mod.co.sngr3* siguen o no una distribución normal, ya que gráficamente (en el gráfico Q-Q), podía observarse que las colas difieren de lo que se consideraría una distribución normal aunque esto podría deberse a los valores *outliers* previamente observados la figura anterior. Para comprobar la normalidad, se aplica la función *Shapiro.test* del paquete *MASS* que hace referencia al test de Shapiro-Wilk. Este test, asume en su hipótesis nula que los residuos siguen una distribución normal. Tras aplicar el test sobre los residuos del modelo *mod.co.sngr3*, se

obtiene un valor de $p=0.11$, es decir, no existe evidencia suficiente para rechazar la hipótesis nula del test *Shapiro-Wilk* y por ello se asume que los residuos del modelo están distribuidos de forma normal.

- **Homocedasticidad/heterocedasticidad:**

Se analiza la homocedasticidad/heterocedasticidad del modelo utilizando el test *Non-Constant Variance Score Test (ncVs)* y el test Breusch-Pagan. Ambos tests asumen en su hipótesis nula que la varianza de los residuos es constante (es decir, existe homocedasticidad) y en la hipótesis alternativa que la varianza cambia según los valores ajustados o la combinación lineal de las variables predictoras, es decir, existe heterocedasticidad. Tras aplicar ambos tests, en ambos se obtienen p-valores superiores al 5%, y por lo tanto se acepta que la varianza de los residuos del modelo planteado es constante (homocedástico).

- **Autocorrelación:**

Para analizar la autocorrelación de los residuos del modelo, se ha utilizado el test de *Durbin-Watson*, que su hipótesis nula define la no autocorrelación (infiriendo independencia) entre los residuos y la alternativa determina que sí existe correlación. Para aplicar este test, es necesario verificar que los residuos se distribuyen de forma normal, lo cual se ha comprobado anteriormente y por lo tanto sí que es posible aplicar el test mediante la función *durbinWatsonTest* sobre el modelo. Del test se obtiene un p-valor = 0.494, y por lo tanto se asume la independencia entre los residuos del modelo, ya que no hay evidencia suficiente para rechazar la hipótesis nula.

- **Multicolinealidad:**

Finalmente, para el análisis de la multicolinealidad, se ha analizado el valor del *Klein* obtenido en el test de Farrar - Glauber, y al igualarse todos los valores de las variables predictoras a cero, se ha asumido que no se ha detectado multicolinealidad entre los residuos del modelo *mod.co.sngr3*. Además, también se ha aplicado la función *vif* - *Variance inflation factor* para cuantificar la correlación entre las variables predictoras del modelo. Como los valores obtenidos para todas las variables predictoras del modelo son cercanos a uno, esto es suficiente para rechazar el principio de multicolinealidad en los residuos del modelo planteado.

Conclusión modelo y comparación

El modelo *mod.co.sngr3* es el único modelo planteado para el cortisol (utilizando la base de datos de la sangre) que cumple con los supuestos cuantificables para un modelo lineal mediante un test, ya que la suposición de linealidad observada en el gráfico de los residuos no es adecuada a simple vista. Es el modelo que más variables predictoras significativas tiene en comparación con los modelos planteados en el Anexo D. Aunque los modelos descartados hayan incumplido algunas suposiciones de los residuos, para comprobar que el *mod.co.sngr3* es efectivamente el modelo con mejores resultados para predecir el nivel de *co.post*, se han aplicado los métodos AIC y BIC y entre todas las combinaciones posibles, es con el que se han obtenido valores más bajos, lo cual es el objetivo que se busca al realizar la comparación de modelos de regresión. La ecuación del modelo *mod.co.sngr3* obtenida es la siguiente:

$$\log(Y) = 6.745 + 0.00039 X_1 + 0.00493 X_2 + 0.00539 X_3 - 0.00024 X_4 + \epsilon$$

Ecuación 5: ecuación final incluyendo los coeficientes de cada covariable para describir el modelo *mod.co.sngr3* y predecir el nivel de cortisol tras aplicar un estímulo sobre el participante, utilizando el conjunto de datos del cortisol para las mediciones obtenidas en la sangre. Transformación logarítmica de la variable respuesta *co.post*.

Siendo cada término,

- $\log(Y)$: variable respuesta *co.post* transformada logarítmicamente.
- 6.745: constante del modelo (B_0)
- X_1 : variable predictora *co.pre*.
- X_2 : variable predictora *age*.
- X_3 : variable predictora *co.reac*.
- X_4 : variable predictora *med.dos*.

2.4.5.2.2 Saliva

Para generar un modelo utilizando únicamente las observaciones de la saliva, lo primero ha sido generar una nueva base de datos, denominada *data.co.slv*, compuesta por 8 variables y 52 observaciones. En comparación con la base de datos principal para el cortisol (*data.co*), se han eliminado cinco variables: *gender* (en el estudio de la saliva son todos hombres, por lo tanto hay un único nivel), *co.meas* (todos se han analizado en la saliva), *disease* (ninguno de los participantes presenta una enfermedad), *med.type* (ninguno toma medicación) y *med.dos* (al no tomar medicación, tampoco debemos mantener la variable que mide la dosis de medicación). Como ya se ha comentado, a cada participante de este estudio se le han aplicado dos tipos de estímulos distintos, por lo que cada *id* de participante se repite dos veces (la variable *id* tendrá la mitad de niveles que participantes/observaciones hay en el conjunto de datos de la saliva), y por lo tanto, la variable edad también se repite para cada uno de ellos en la observación de cada tipo de estímulo. Se ha observado que únicamente existe un 0.01% de observaciones faltantes en el conjunto de datos general, ya que falta la medición de *co.pre* (nivel de cortisol previo al estímulo) en un paciente, y por lo tanto también se obtiene un valor faltante en las variables *co.reac* y *co.res*, las cuales se generan a raíz de los valores medidos de cortisol. Aunque la distribución de los datos no variará mucho del análisis exploratorio llevado a cabo en los subapartados anteriores para los conjuntos de datos con una cantidad de observaciones y variables diferentes, dado que el número de observaciones ha disminuido, a continuación se vuelve a mostrar un análisis de esas variables. Finalmente, también se volverá a analizar la correlación entre variables, ya que el coeficiente de correlación entre las variables sí que cambiará al haber modificado el conjunto de datos.

En la Figura 30 se muestra un gráfico de cajas de la variable respuesta *co.post* (nivel de cortisol tras aplicar el estímulo) en este conjunto de datos (se observa un valor *outlier* en la parte superior que coincide con el valor máximo de la variable) y en la Tabla 22 se puede observar un resumen numérico de la variable donde se recoge el valor mínimo, el máximo, la media junto a la desviación estándar, la mediana y el primer y tercer cuantil.

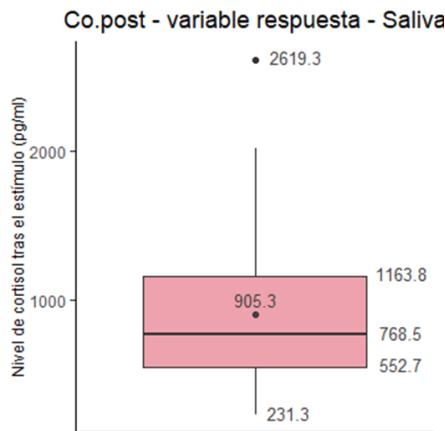


Figura 30: boxplot de la variable cortisol tras aplicar un estímulo sobre el participante, donde se muestran los valores de la media, mediana Q1, Q3, min y max, utilizando el conjunto de datos con las mediciones en la saliva

Tabla 22: descriptiva numérica de la variable respuesta co.post (nivel de cortisol tras aplicar un estímulo sobre el participante) para el conjunto de datos con mediciones en la saliva

	Co.post
Valor general	
Min	231.26
Q1	552.74
Mediana	768.5
Media (SD)	905.3 (508.63)
Varianza	258704.5
Q3	1163.81
Max	2619.29
Rango	2388.03
IQR	611.07

Para la variable respuesta *co.post* en el conjunto de datos de la saliva, no hay ningún valor faltante. Respecto a la distribución de la variable para el conjunto de datos reducido, se aplica el test de Shapiro-Wilk mediante la función *normality()* del paquete *dlookr*, y se obtiene un p-valor inferior al 5% (p-valor=0.001), por lo tanto no se acepta la hipótesis nula y no se considera que la variable respuesta *co.post* siga una distribución normal. De forma gráfica, esto se analiza en la Figura 31, donde se observa que la variable está sesgada a la derecha cuando no se le aplica ninguna transformación. Sin embargo, parece que a simple vista la distribución mejora cuando se le aplica una transformación logarítmica, y esto se corrobora con el test de Shapiro-Wilk sobre la variable transformada, donde se obtiene un p-valor = 0.966, muy alto y por lo tanto aceptando la hipótesis nula de normalidad.

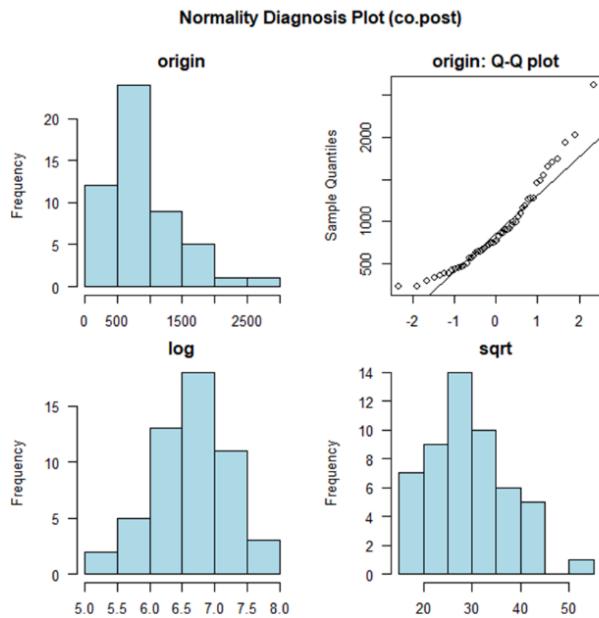


Figura 31: distribución de la variable respuesta que mide el nivel de cortisol tras aplicar un estímulo sobre el participante (co.post). Arriba a la izquierda, histograma de la distribución original. Arriba a la derecha, gráfico QQ de los datos originales. Los gráficos de abajo muestran histogramas de la distribución de la variable en caso de aplicar la transformación logarítmica o de raíz cuadrada a los datos. Conjunto de datos del cortisol con mediciones de la saliva

Respecto a las variables predictoras, en la siguiente Figura 32 se muestra la distribución de las mismas:

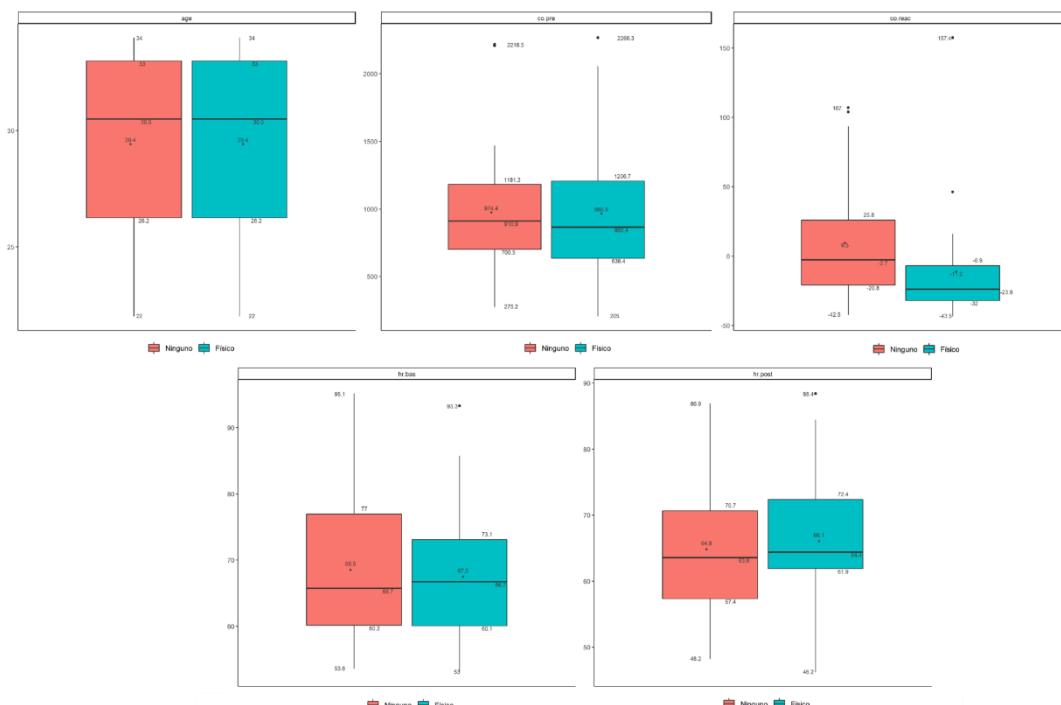


Figura 32: boxplots con los valores de la media, mediana, Q1, Q3, min y max para las variables numéricas del conjunto de datos con las mediciones de la saliva según el estímulo aplicado. Fila de arriba: variable edad, nivel de cortisol previo y reacción del cortisol. Fila de abajo: niveles del ritmo cardiaco (hr.bas y hr.post).

En la Tabla 23 se resumen los datos más significativos de cada una de las variables para este conjunto de datos. Los datos se muestran de manera general, puesto que en la Tabla 14 y Tabla 15 mostradas anteriormente ya se ha especificado el EDA para cada uno de los tipos de estímulos.

Tabla 23: descriptiva numérica de las covariables co.pre, co.reac, age y ritmos cardiacos de forma general. Se recogen valores generales (min, max, media, mediana, Q1, Q3) y valores de las medidas de dispersión de cada una (varianza, rango, IQR). Conjunto de datos del cortisol con mediciones de saliva

	Variable				
	Co.pre	Co.reac	Age	Hr.bas	Hr.post
Valor general					
Min	205.0	-43.52	22.00	53.05	46.19
Q1	641.53	-28.15	26.00	60.07	58.74
Median	910.93	-17.24	30.50	66.01	63.71
Media (SD)	970.04 (490.91)	-1.05 (42.75)	29.42 (4.07)	68.00 (10.31)	65.44 (9.47)
Varianza	240992.6	1827.56	16.56	106.30	89.68
Q3	1183.86	14.30	33.00	74.46	71.08
Max	2266.34	157.44	34.00	95.13	88.41
Rango	2061.34	200.96	12.00	42.08	42.22
IQR	542.33	42.45	7.00	14.39	12.33

La distribución de las variables *hr.post* y *hr.bas* es la misma en este conjunto de datos que en el conjunto de datos para el cortisol general (se puede observar en la Figura 20 puesto que únicamente teníamos observaciones de estas variables en las muestras obtenidas mediante la saliva). La distribución de las variables *co.reac* (*índice de reacción al cortisol*), *age*, y *co.pre* (nivel de cortisol previo al estímulo) ha variado respecto al conjunto de datos original (Figura 33), pero en ninguno de los casos esto ha hecho que la distribución de la variable se asemeje a la normal, puesto que se obtienen p-valores inferiores al 5%, y por lo tanto no se puede aceptar la hipótesis nula (a excepción de *hr.post*, tal y como se había comentado para el conjunto de datos general). Al transformar las variables logarítmicamente, todas las variables excepto *age* son significativas al 5%, por lo tanto sí que se aceptaría la hipótesis de normalidad para las variables *hr.bas*, *co.pre*, *co.reac* y *hr.post* en este conjunto de datos reducido.

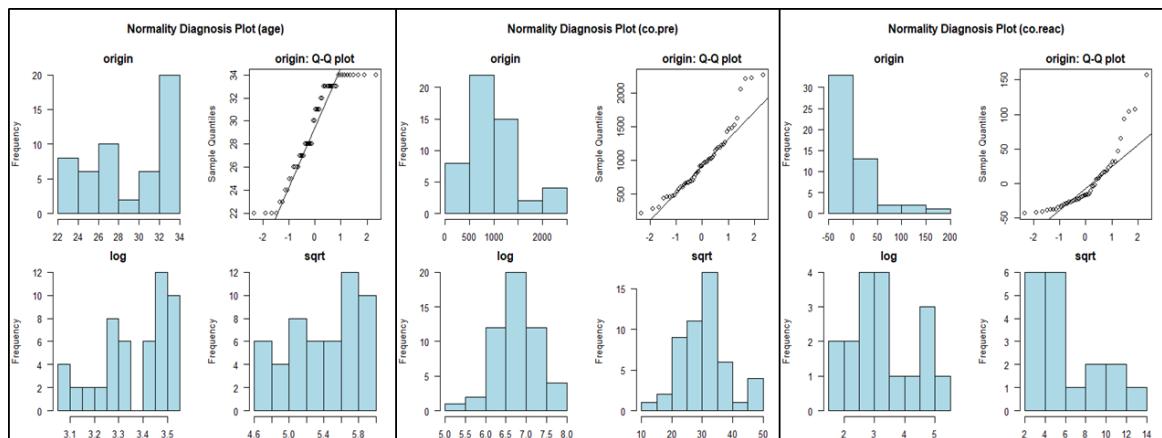


Figura 33: análisis de la normalidad. De izq. a dcha: variables age, co.pre y co.reac. Conjunto de datos del cortisol medido en la saliva. Para cada variable se muestra la distribución original mediante histograma y gráfico QQ, e histograma con transformación log y sqrt

Finalmente, respecto a las correlaciones entre las variables, a continuación se muestra el mapa de calor (heatmap, Figura 34) obtenido a partir del conjunto de datos y la matriz de correlaciones (Tabla 24). En la tabla se muestran los valores de los coeficientes de correlación para este caso. Se observa que los coeficientes para los ritmos cardiacos tienen el mismo valor (0.862, correlación muy fuerte y positiva) que en el conjunto de datos general, puesto que las mediciones de la sangre no tenían influencia sobre ellas. La correlación entre *co.res* y *co.reac*

sigue siendo alta (ya que *co.res* se genera a partir de *co.reac*) y también la relación entre el cortisol previo y el posterior es bastante alta y positiva (0.726), siendo algo menor que para el conjunto de datos general.

Tabla 24: matriz de correlaciones para las variables del conjunto de datos del cortisol medido en la saliva

	age	stimulus.type	Co.pre	Co.post	Co.reac	Co.res	hr.bas	hr.post
age	1							
stimulus.type	0	1						
co.pre	0.16	-0.03	1					
co.post	0.14	-0.18	0.73	1				
Co.reac	-0.06	-0.33	-0.28	0.39	1			
Co.res	-0.08	-0.24	0.05	0.63	0.75	1		
hr.bas	0.34	-0.04	0.22	-0.01	-0.28	-0.30	1	
hr.post	0.42	0.10	0.14	-0.06	-0.26	-0.29	0.86	1

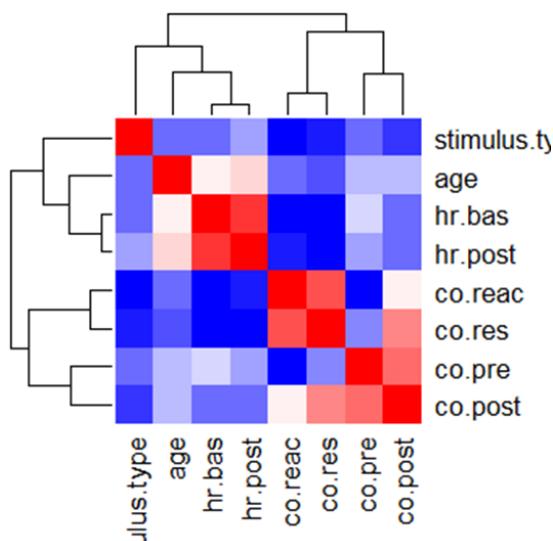


Figura 34: mapa de calor (heatmap) a partir de las correlaciones para las variables del conjunto de datos del cortisol medido en la saliva. Los rectángulos rojos identifican los coeficientes de correlación más cercanos a uno (más intensidad de rojo mayor correlación), y los rectángulos azules, menor correlación (mayor intensidad de azul menor correlación).

Una vez resumidas las variables de este conjunto de datos, se procede a explicar el modelo generado las variables.

Modelo saliva - cortisol

En la tabla de correlaciones (Tabla 24) y en la Figura 34 anterior se ha mostrado que las variables por pares con un coeficiente de correlación más alto son *hr.bas* y *hr.post*, seguidas por *co.reac* y *co.res*. A la hora de diseñar el modelo, no será posible incluir las cuatro variables como variables predictoras, ya que se incumpliría la condición de independencia entre ellas. Por lo tanto, en el caso del par *hr.bas*-*hr.post*, se escoge incluir en el modelo *hr.post*. La variable *hr.post* muestra una correlación ligeramente más alta que *hr.bas* con la variable respuesta (lo que es deseable), y su correlación frente a la variable *co.pre* (variable que indudablemente debe estar en el modelo) es más baja que la de *hr.bas*. En relación a las variables *co.reac* y *co.res*, se mantiene la variable *co.reac* por tratarse de una variable numérica y no una variable categórica, aunque su correlación con *co.pre* sea ligeramente superior y con la variable respuesta ligeramente inferior (esta diferencia no se ha considerado significativa).

El modelo escogido para predecir el nivel de cortisol utilizando la base de datos de la saliva se denomina *mod.co.slv2*, y en este modelo se han transformado todas las variables numéricas en logarítmicas para mejorar la normalidad de los residuos del modelo. En comparación con los otros tres modelos que se han generado, es el modelo con el que mejores resultados se han obtenido, y al hacer la comparación con los otros (mostrados en el Anexo E del documento), es con el que se han obtenido valores más bajos para las funciones de AIC y BIC. Ninguno de los otros modelos ha cumplido el supuesto de la normalidad en los residuos, y únicamente uno de ellos ha mostrado homocedasticidad en los residuos, con un p-valor superior al 5% (modelo en el que se ha transformado logarítmicamente únicamente la variable respuesta). Por ello, los tres modelos presentados en el Anexo quedan descartados para predecir el nivel de cortisol en saliva al haber aplicado un estímulo sobre el paciente.

En el planteamiento inicial del modelo *mod.co.slv2*, éste estaba compuesto por las variables numéricas *co.pre* (*nivel de cortisol previo*), *age*, *co.reac* (*índice de reacción del cortisol*) y ritmo cardíaco post estímulo, *hr.post* (todas ellas transformadas logarítmicamente) y la variable predictora categórica que define el tipo de estímulo. Sin embargo, únicamente las variables *log(co.pre)* y *log(co.reac)* han resultado ser significativas al 5% para predecir la variable respuesta *log(co.post)*, por lo tanto se ha aplicado Akaike (mediante la función *stepAIC*) para determinar si efectivamente se debían eliminar las demás variables del modelo. Finalmente, el modelo con doble transformación logarítmica que se ha planteado ha sido el siguiente:

$$\log(Y) = B_0 + B_1 \log(X_{co.pre}) + B_2 \log(X_{co.reac}) + \epsilon$$

*Ecuación 6: planteamiento inicial del modelo mod.co.slv2 utilizando el conjunto de datos del cortisol para las mediciones en saliva y predecir el nivel de cortisol tras la aplicación de un estímulo sobre el participante.
Transformación logarítmica de la variable respuesta y las covariables.*

En la Tabla 25 se muestra el *output* obtenido del modelo:

Tabla 25: resultados del modelo de regresión, logarítmico de las variables co.pre y co.reac como predictores del nivel de cortisol tras la aplicación del estímulo, también transformado logarítmicamente

Predictores	Coeficiente B	Std.Err	t	Sig
Constante	-0.27953	0.13963	-2.002	0.0684
log(co.pre)	0.94903	0.02224	42.672	1.78e-14 ***
log(co.reac)	0.27674	0.01635	16.925	9.70e-10 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1				
F	1472			
R ²	0.9953			
p-valor	4.485e-15			

En la Tabla 25, se observa que finalmente el modelo está compuesto por las variables *log(co.pre)* y *log(co.reac)*, ambas significativas y con el valor ajustado *R*² del modelo muy alto. El p-valor del modelo también es muy significativo. En la siguiente imagen se muestra el comportamiento de los residuos del modelo definido.

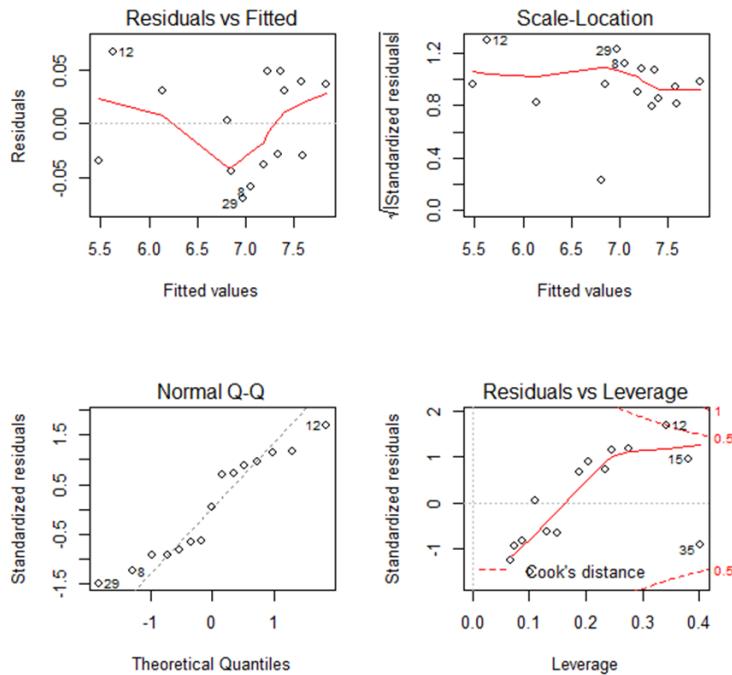


Figura 35: distribución de los residuos del modelo mod.co.slv2 (transformación logarítmica de la variable respuesta y las covariables). Gráfico de linealidad (arriba izq.), homocedasticidad (arriba dcha.), normalidad (abajo izq.) y puntos outliers o influyentes (abajo dcha.)

En la Figura 35, se muestra el comportamiento de los residuos del modelo, en términos de normalidad, homocedasticidad, valores *outliers* y linealidad. En términos de la linealidad, ésta no se cumple puesto que la línea roja muestra un pico hacia abajo en el gráfico, por lo que no parece que se cumpla la suposición de la relación lineal entre los residuos. Cabe destacar, que del conjunto de datos original se han eliminado tres valores influentes (*outliers*, en concreto las observaciones número 33, 46 y 7), ya que no se cumplía la hipótesis de normalidad con la influencia de estas tres observaciones. Tras eliminarlos, han surgido nuevos valores que se encuentran separados del resto, pero no se considera que se deban eliminar ya que las suposiciones del modelo seguirían siendo similares y por lo tanto su influencia no es tan alta.

- **Normalidad de los residuos:**

Respecto al análisis de los residuos, se ha aplicado el test de Shapiro-Wilk sobre ellos para analizar la distribución normal, y se ha obtenido un *p*-valor= 0.1246, por lo tanto no hay evidencia suficiente para rechazar la hipótesis nula de normalidad de los datos. En la Figura 35 (gráfico *Normal QQ*) no parece que a simple vista las observaciones sigan una distribución normal, y esto parece estar influenciado por las pocas observaciones del conjunto de datos, la cual está compuesta por 52 observaciones tras haber eliminado los tres valores influentes mencionados previamente. Sin embargo, como se ha obtenido un *p*-valor superior al 5%, sí que se acepta que los residuos del modelo se distribuyen de manera normal.

- **Homocedasticidad/ heterocedasticidad:**

Respecto a la homocedasticidad de los residuos, gráficamente es complicado determinar cómo es la varianza, ya que la línea roja del gráfico *Scale-Location* no es horizontal y parece que una vez más es debido al reducido tamaño del conjunto de datos. Al aplicar sobre los datos el *ncVs* test y el test *Breusch-Pagan*, se ha obtenido en ambos *p*-valores superiores a 0.05, por lo tanto no existe evidencia suficiente para rechazar la hipótesis nula y se asume que la varianza de los residuos es constante.

- **Autocorrelación:**

Para analizar la autocorrelación de los residuos del modelo, se ha aplicado el test de *Durbin-Watson*, el cual en su hipótesis nula define la independencia entre los residuos. Para aplicar el test de autocorrelación, se ha comprobado previamente que los residuos siguen una distribución normal. Finalmente, se ha obtenido un p-valor= 0.34, y por lo tanto se acepta la independencia entre los residuos del modelo.

- **Multicolinealidad:**

Finalmente, para el análisis de la multicolinealidad, se ha analizado una vez más el valor del *Klein* obtenido en el test de Farrar - Glauber, y los valores del *klein* para $\log(co.\text{pre})$ y $\log(co.\text{reac})$ son nulos (igualados a cero), por lo tanto se ha asumido que no se ha detectado multicolinealidad entre los residuos del modelo. Además, también se ha aplicado la función *vif - Variance inflation factor* para cuantificar la correlación entre las variables predictoras del modelo, y los valores obtenidos para ambas variables son cercanos a uno, por lo tanto, suficiente para rechazar el principio de multicolinealidad en los residuos del modelo analizado.

Conclusión modelo y comparación

El modelo *mod.co.slv2* es el modelo que utilizando la base de datos de la saliva mejores resultados ha proporcionado, en comparación con los que se presentan en el Anexo E de este documento. Aunque la linealidad de los modelos no parece que se cumpla al analizar el conjunto de datos, se han obtenido los valores más bajos para las funciones AIC y BIC (método Akaike) para la selección de modelos. La ecuación del modelo *mod.co.slv2* con los coeficientes de cada variable es la siguiente:

$$\log(Y) = -0.280 + 0.949 \log(X_1) + 0.277 \log(X_2) + \epsilon$$

Ecuación 7: ecuación final incluyendo los coeficientes de cada covariable para describir el modelo mod.co.slv2 y predecir el nivel de cortisol tras aplicar un estímulo sobre el participante, utilizando el conjunto de datos del cortisol para las mediciones obtenidas de la saliva. Transformación de la variable respuesta co.post y las covariables.

Siendo cada término,

- $\log(Y)$: variable respuesta *co.post* transformada logarítmicamente.
- -0.280: constante del modelo (B_0)
- X_1 : variable predictora *co.pre* transformada logarítmicamente.
- X_2 : variable predictora *co.reac* transformada logarítmicamente.

2.4.6 Conclusión modelo cortisol

Por lo tanto, una vez analizado los modelos del cortisol utilizando la base de datos completa *data.co* (propuesta 1) se ha observado que no se han cumplido las suposiciones para un modelo lineal. Al separar el conjunto de datos por tipos de medición del cortisol (propuesta 2), se ha observado que en los modelos planteados se han cumplido los supuestos de los residuos, a excepción de la linealidad. Este supuesto se ha analizado gráficamente en ambos casos (Figura 29 y Figura 35), y se ha observado que en ninguno de ellos se ha cumplido una relación lineal en los residuos. Al comparar el modelo de la sangre con el de la saliva, se ha observado que el modelo de la sangre muestra un valor ligeramente inferior (para las funciones AIC y BIC), por lo tanto se podría decir que se ajusta mejor a los datos que el modelo de la saliva, aunque esto podría ser debido a que tiene más observaciones que el conjunto de datos de la saliva. Sin embargo, se espera poder aplicar ambos modelos en un conjunto de datos más grande en cada caso, ya que se cree que la suposición de linealidad no se cumple en gran parte, debido al pequeño tamaño de la muestra.

2.5 Aplicación de los modelos

Uno de los objetivos principales del trabajo es analizar si la variable etnia es significativa para el estrés medido mediante los niveles de oxitocina y cortisol tras aplicar diferentes estímulos sobre los participantes. Tal y como se ha explicado en los apartados anteriores, los modelos se han definido a partir de datos de estudios previamente publicados en los cuales la variable etnia no estaba incluida, y entonces tampoco lo está en ninguna de las regresiones planteadas para cada uno de los modelos.

En un principio se planteó que los modelos generados se podrían aplicar sobre el conjunto de datos perteneciente al estudio piloto de la Universidad de Maryland, añadiendo como covariable la etnia, dato que sí que se recoge de los participantes en el estudio. Sin embargo, el desarrollo del trabajo ha hecho que algunas variables (como el ritmo cardiaco por ejemplo) resulten significativas para mostrar los cambios en ambos biomarcadores, y éstas no se han tenido en cuenta a la hora de recopilar los datos de los participantes en el estudio comenzado en 2018. Además, debido a la pandemia SARS-CoV-2/COVID-19 actual y al colapso que la situación ha generado en laboratorios de todo el mundo, únicamente se ha llevado a cabo el análisis de las muestras de la oxitocina. Por todo ello, no se han podido aplicar los modelos sobre el conjunto de datos del estudio piloto.

Sin embargo, se ha querido mostrar cual sería el procedimiento a seguir para responder a la pregunta de investigación cuando se pueda disponer de los datos y se mejore el protocolo actual de la recogida de muestras para el desarrollo junto con la Universidad de Maryland. Se ha simulado la variable etnia utilizando los datos oficiales presentados por el censo de Estados Unidos para el estado de Maryland (*United States Census Bureau*⁵) en el año 2019. El censo define que la población del estado de Maryland está distribuida de la siguiente manera: 50% blancos, 29.8% afroamericanos, 10.70% hispanos o latinos y 9.50% pertenecientes a otra etnia (donde se incluyen por ejemplo los indios americanos o nativos de Alaska, asiáticos, hawaianos o isleños del Pacífico). Para cada modelo presentado, se ha añadido como covariable la variable categórica nominal etnia, definida con los cuatro niveles mencionados. Las etnias se han aplicado sobre el conjunto de datos de manera aleatoria, asignando a las 84 observaciones del conjunto de datos inicial 42 personas blancas, 25 afroamericanas, 9 hispanas y 8 personas pertenecientes a la categoría restante.

El procedimiento llevado a cabo para analizar si la variable es o no significativa se ha añadido en el Anexo F del presente documento. En él se resumen los resultados obtenidos tras aplicar en la ecuación de cada biomarcador (oxitocina general, cortisol en las medidas de saliva y sangre) la covariable etnia (sin aplicar ninguna transformación sobre ella al tratarse de una variable categórica). Para cada uno de los modelos, se ha analizado si alguno de los niveles de la variable predictora etnia es significativa al 5% y también, si al añadir la variable, el modelo sufría alguna modificación (alguna variable que se había considerado significativa dejaba de serlo, el ajuste del modelo había empeorado etc.).

Como era de esperar, al haber incluido los valores de la etnia de manera aleatoria, en ninguno de los casos ésta ha resultado ser significativa, y los modelos tampoco se han modificado en relación al R² y al p-valor. Todas las variables que previamente se incluían en el modelo han seguido siendo significativas aunque se haya añadido la variable mencionada.

⁵ <https://www.census.gov/quickfacts/MD>

Los resultados obtenidos no responden a la pregunta de investigación planteada en el presente proyecto, y no es posible determinar si la etnia tiene una influencia o no sobre los valores de oxitocina y cortisol. Sin embargo, se ha conseguido plasmar cuál debería ser el procedimiento adecuado para analizar la variable etnia sobre el conjunto de datos del proyecto piloto una vez se realicen las mediciones de ambos biomarcadores, la muestra sea más grande y se disponga de los datos de todas las covariables que se han considerado relevantes a lo largo del presente trabajo.

2.6 Repositorio online

Los archivos generados durante el desarrollo del presente proyecto son accesibles a través de Github en el siguiente enlace: <https://github.com/jonerenteria/TFM>.

3. Conclusiones

Uno de los objetivos principales ha sido plantear un modelo de regresión utilizando los biomarcadores de la oxitocina y el cortisol como variables dependientes para analizar los factores que afectan al estrés en un individuo. Para cumplir el objetivo, se ha realizado una búsqueda exhaustiva de estudios previamente publicados y relacionados con la temática para generar un conjunto de datos y utilizarlo en el análisis de los modelos. La combinación de las covariables con mejor respuesta ha definido cada uno de los modelos de regresión descritos en la memoria. Además, este trabajo evidencia la necesidad de revisar las variables definidas en el proyecto en el que se basa este estudio y actualizar el protocolo actual de las visitas a los hogares para la recogida de datos. La literatura y el análisis realizado han demostrado que por ejemplo el ritmo cardiaco es una variable importante a la hora de analizar el estrés en una persona. Por ello, cuando el proyecto vuelva a activarse, también se recogerán los datos del ritmo cardiaco previo y posterior al estímulo de los participantes. Cabe destacar que aparte de los datos puramente demográficos (como la edad o etnia) del mismo modo en el proyecto se recogerán datos relacionados con la salud (altura, peso, ingesta de alcohol etc.) y también aquellos asociados con aspectos socio-psicológicos como la relación en pareja, la depresión o los hábitos diarios. Una vez recogidos todos los datos, se podrán plantear regresiones alternativas utilizando variables que no se han considerado en los estudios de la literatura.

Otro de los objetivos es analizar el efecto de la etnia para el nivel de estrés de una persona. Sin embargo, este objetivo se ha visto afectado por la actual situación del SARS-CoV-2/COVID-19. En un principio, se pretendían aplicar los modelos de regresión sobre el conjunto de datos perteneciente a la Universidad de Maryland para analizar el efecto de la etnia. Además, se valoraría la posibilidad de modificar alguna de las covariables incluidas (o incluir nuevas) en caso necesario. No obstante, el colapso en los laboratorios ha hecho que uno de los biomarcadores necesarios para definir el estrés no haya podido ser analizado, impidiendo que el modelo pudiera ser aplicado sobre el conjunto de datos del proyecto.

El desarrollo de este trabajo ha sido muy satisfactorio personalmente, ya que me ha permitido conocer dos caras de la investigación. Por un lado, el análisis teórico basado en datos de la literatura me ha brindado la oportunidad de aplicar métodos estadísticos directamente relacionados con el máster cursado sobre un conjunto de datos preparado para realizar técnicas de regresión. Por otro lado, el trabajar con datos reales, me ha ayudado a conocer todo el proceso de análisis, desde la recogida de datos que necesita un protocolo adecuado hasta la interpretación de los resultados.

Por último, aunque he podido desarrollar teóricamente el procedimiento a aplicar para dar respuesta a la pregunta principal de investigación, espero que cuando la situación de la pandemia vuelva a la normalidad, se pueda seguir con esta línea de investigación junto con la Universidad de Maryland y poder completar el trabajo en un futuro cercano.

4. Bibliografía

- Alley, Jenna, Lisa M Diamond, David L Lipschitz, y Karen Grewen. 2019. «Associations between oxytocin and cortisol reactivity and recovery in response to psychological stress and sexual arousal». *Psychoneuroendocrinology* 106: 47-56.
- Anderson, Norman B. 1998. «Levels of Analysis in Health Science: A Framework for Integrating Sociobehavioral and Biomedical Research». *Annals of the New York Academy of Sciences* 840 (1): 563-76. <https://doi.org/10.1111/j.1749-6632.1998.tb09595.x>.
- Anderson, Norman B, Rodolfo A Bulatao, Barney Cohen, Panel on Race, y National Research Council. 2004. «Cumulative psychosocial risks and resilience: A conceptual perspective on ethnic health disparities in late life». En *Critical perspectives on racial and ethnic differences in health in late life*. National Academies Press (US).
- Arias, Adalberto Campo, Heidi Oviedo, y Edwin Herazo. 2015. «Escala de Discriminación en la Vida Cotidiana: Consistencia y estructura interna en estudiantes de medicina». *Revista Médica de Risaralda* 21 (2): 1.
- Barrera, Mónica Alejandra Mondragón. 2014. «Uso de la correlación de Spearman en un estudio de intervención en fisioterapia». *Movimiento Científico* 8 (1): 98-104.
- Bennett, Gary G., Marcellus M. Merritt, y Kathleen Y. Wolin. 2004. «Ethnicity, education, and the cortisol response to awakening: A preliminary investigation». *Ethnicity & Health* 9 (4): 337-47. <https://doi.org/10.1080/1355785042000285366>.
- Bischoff, M., V. Howland, J. Klinger-König, S. Tomczyk, S. Schmidt, M. Zygmunt, M. Heckmann, et al. 2019. «Save the children by treating their mothers (PriVileG-M-study) - study protocol: a sequentially randomized controlled trial of individualized psychotherapy and telemedicine to reduce mental stress in pregnant women and young mothers and to improve Child's health». *BMC Psychiatry* 19 (1): 371. <https://doi.org/10.1186/s12888-019-2279-0>.
- Boileau, Kayla, Kheana Barbeau, Rupali Sharma, y Catherine Bielajew. 2019. «Ethnic Differences in Diurnal Cortisol Profiles in Healthy Adults: A Meta-Analysis». *British Journal of Health Psychology* 24 (4): 806-27. <https://doi.org/10.1111/bjhp.12380>.
- Cabrera, Natasha, Lina Guzman, Kimberly Turner, Jenessa Malin, y P Mae Cooper. 2016. «A national portrait of the health and education of Hispanic boys and young men».
- Cardoso, Christopher, Mark A Ellenbogen, Mark Anthony Orlando, Simon L Bacon, y Ridha Joober. 2013. «Intranasal oxytocin attenuates the cortisol response to physical stress: a dose-response study». *Psychoneuroendocrinology* 38 (3): 399-407.
- Coleman, Karen J, Christine Stewart, Beth E Waitzfelder, John E Zeber, Leo S Morales, Ameena T Ahmed, Brian K Ahmedani, et al. 2016. «Racial/Ethnic Differences in Diagnoses and Treatment of Mental Health Conditions across Healthcare Systems Participating in the Mental Health Research Network». *Psychiatric services (Washington, D.C.)* 67 (7): 749-57. <https://doi.org/10.1176/appi.ps.201500217>.
- Estrada-Y-Martin, Rosa M., y Philip R. Orlander. 2011. «Salivary Cortisol Can Replace Free Serum Cortisol Measurements in Patients With Septic Shock». *Chest* 140 (5): 1216-22. <https://doi.org/10.1378/chest.11-0448>.
- Gallo, Linda C., Frank J. Penedo, Karla Espinosa de los Monteros, y William Arguelles. 2009. «Resiliency in the Face of Disadvantage: Do Hispanic Cultural Characteristics Protect Health Outcomes?». *Journal of Personality* 77 (6): 1707-46. <https://doi.org/10.1111/j.1467-6494.2009.00598.x>.
- Goosby, Bridget J, y Chelsea Heidbrink. 2013. «The transgenerational consequences of discrimination on African-American health outcomes». *Sociology compass* 7 (8): 630-43.
- Halloran, Michael J. 2019. «African American Health and Posttraumatic Slave Syndrome: A Terror Management Theory Account». *Journal of Black Studies* 50 (1): 45-65. <https://doi.org/10.1177/0021934718803737>.

- Hammond, G. L., C. L. Smith, y D. A. Underhill. 1991. «Molecular Studies of Corticosteroid Binding Globulin Structure, Biosynthesis and Function». *The Journal of Steroid Biochemistry and Molecular Biology* 40 (4): 755-62. [https://doi.org/10.1016/0960-0760\(91\)90300-T](https://doi.org/10.1016/0960-0760(91)90300-T).
- Heinrichs, Markus, Thomas Baumgartner, Clemens Kirschbaum, y Ulrike Ehlert. 2003. «Social support and oxytocin interact to suppress cortisol and subjective responses to psychosocial stress». *Biological psychiatry* 54 (12): 1389-98.
- Hwang, Wei-Chin, y Julia Y. Ting. 2008. «Disaggregating the effects of acculturation and acculturative stress on the mental health of Asian Americans». *Cultural Diversity and Ethnic Minority Psychology* 14 (2): 147-54. <https://doi.org/10.1037/1099-9809.14.2.147>.
- Irizar, Karmele Salaberria, y Analia del Valle Sanchez Haro. 2017. «Estrés migratorio y salud mental». *Psicología Conductual* 25 (2): 419.
- Juster, Robert-Paul, Bruce S. McEwen, y Sonia J. Lupien. 2010. «Allostatic Load Biomarkers of Chronic Stress and Impact on Health and Cognition». *Neuroscience & Biobehavioral Reviews, Psychophysiological Biomarkers of Health*, 35 (1): 2-16. <https://doi.org/10.1016/j.neubiorev.2009.10.002>.
- Kaufman, Eliaz and Lamster, Ira B. 2002. «The diagnostic applications of saliva—a review». *Critical Reviews in oral biology & medicine - SAGE Publications* 13 (2): 197-212.
- Kronenberg, G., J. Schöner, C. Nolte, A. Heinz, M. Endres, y Karen Gertz. 2017. «Charting the Perfect Storm: Emerging Biological Interfaces between Stress and Stroke». *European Archives of Psychiatry and Clinical Neuroscience* 267 (6): 487-94. <https://doi.org/10.1007/s00406-017-0794-x>.
- Kubzansky, Laura D, Wendy Berry Mendes, Allison A Appleton, Jason Block, y Gail K Adler. 2012. «A heartfelt response: oxytocin effects on response to social stress in men and women». *Biological psychology* 90 (1): 1-9.
- Kumsta, Robert, y Markus Heinrichs. 2013. «Oxytocin, Stress and Social Behavior: Neurogenetics of the Human Oxytocin System». *Current Opinion in Neurobiology, Neurogenetics*, 23 (1): 11-16. <https://doi.org/10.1016/j.conb.2012.09.004>.
- Lee, Do Yup, Eosu Kim, y Man Ho Choi. 2015. «Technical and clinical aspects of cortisol as a biochemical marker of chronic stress». *BMB Reports* 48 (4): 209-16. <https://doi.org/10.5483/BMBRep.2015.48.4.275>.
- Luecken, Linda J, David P MacKinnon, Shannon L Jewell, Keith A Crnic, y Nancy A Gonzales. 2015. «Effects of prenatal factors and temperament on infant cortisol regulation in low-income Mexican American families». *Developmental psychobiology* 57 (8): 961-73.
- Martínez Ortega, Rosa María. 2009. «El coeficiente de correlación de los rangos de Spearman caracterización». *Revista Habanera de Ciencias Médicas* 8 (2): 0-0.
- McCullough, Michael E., Patricia Smith Churchland, y Armando J. Mendez. 2013. «Problems with Measuring Peripheral Oxytocin: Can the Data on Oxytocin and Human Behavior Be Trusted?». *Neuroscience & Biobehavioral Reviews* 37 (8): 1485-92. <https://doi.org/10.1016/j.neubiorev.2013.04.018>.
- Miller, Robert, Franziska Plessow, Clemens Kirschbaum, y Tobias Stalder. 2013. «Classification Criteria for Distinguishing Cortisol Responders From Nonresponders to Psychosocial Stress: Evaluation of Salivary Cortisol Pulse Detection in Panel Designs». *Psychosomatic Medicine* 75 (9): 832-40. <https://doi.org/10.1097/PSY.0000000000000002>.
- National Center for Health Statistics. 2017. *Health, United States, 2016, with Chartbook on Long-Term Trends in Health*. Government Printing Office.
- Öhman, Lena, Jan Bergdahl, Lars Nyberg, y Lars-Göran Nilsson. 2007. «Longitudinal Analysis of the Relation between Moderate Long-Term Stress and Health». *Stress and Health* 23 (2): 131-38. <https://doi.org/10.1002/stm.1130>.
- Ooishi, Yuuki, Hideo Mukai, Ken Watanabe, Suguru Kawato, y Makio Kashino. 2017. «Increase in salivary oxytocin and decrease in salivary cortisol after listening to relaxing slow-tempo and exciting fast-tempo music». *PloS one* 12 (12): e0189075.

- Panchang, Sarita, Hilary Dowdy, Rachel Kimbro, y Bridget Gorman. 2016. «Self-Rated Health, Gender, and Acculturative Stress among Immigrants in the U.S.: New Roles for Social Support». *International Journal of Intercultural Relations* 55 (noviembre): 120-32. <https://doi.org/10.1016/j.ijintrel.2016.10.001>.
- Peters, JR and Walker, RF and And, D RIAD-FAHMY and Hall, R. 1982. «Salivary cortisol assays for assessing pituitary-adrenal reserve». *Clinical Endocrinology - Wiley Online Library* 17 (6): 583-92.
- Salleh, Mohd. Razali. 2008. «Life Event, Stress and Illness». *The Malaysian Journal of Medical Sciences : MJMS* 15 (4): 9-18.
- Sue Carter, C. 1998. «NEUROENDOCRINE PERSPECTIVES ON SOCIAL ATTACHMENT AND LOVE». *Psychoneuroendocrinology* 23 (8): 779-818. [https://doi.org/10.1016/S0306-4530\(98\)00055-9](https://doi.org/10.1016/S0306-4530(98)00055-9).
- Tas, Cumhur, Elliot C Brown, Gokcer Eskikurt, Sezen Irmak, Orkun Aydin, Aysen Esen-Danaci, y Martin Brüne. 2018. «Cortisol response to stress in schizophrenia: associations with oxytocin, social support and social functioning». *Psychiatry research* 270: 1047-52.

ANEXOS

Índice Anexo

Anexo A: Generación base de datos	IV
Anexo B: Modelo oxitocina con el conjunto de datos completo	VII
Modelo I.....	VII
Modelo II.....	IX
Modelo III.....	XI
Anexo C: Modelo cortisol con el conjunto de datos completo	XV
Modelo I.....	XV
Modelo II.....	XV
Modelo III.....	XVI
Anexo D: Modelo cortisol con el conjunto de datos con mediciones en sangre	XVII
Modelo I.....	XVII
Modelo II.....	XVIII
Modelo III.....	XIX
Anexo E: Modelo cortisol con el conjunto de datos con mediciones en saliva	XXI
Modelo I.....	XXI
Modelo II.....	XXI
Modelo III.....	XXII
Anexo F: Aplicación de los modelos.....	XXIII

Lista de figuras

Figura I.B: residuos del primer modelo planteado mod.oxt para predecir el nivel de oxitocina tras aplicar un estímulo. Gráfico de linealidad (arriba izq.), homocedasticidad (arriba dcha.), normalidad (abajo izq.) y puntos outliers o influyentes (abajo dcha.)	VIII
Figura II.B: análisis de los residuos en términos de linealidad, homocedasticidad, normalidad y valores influyentes para el modelo mod.oxt3	X
Figura III.B: estimación del valor de lambda para el modelo que predice el nivel de oxitocina post aplicación de un estímulo sin ninguna transformación (mod.oxt). En el gráfico de la izquierda, se muestra el intervalo de confianza para el valor de lambda, y en la derecha se observa que el valor es cercano a 0.20.....	XII
Figura IV.B: análisis de los residuos en términos de linealidad, homocedasticidad, normalidad y valores influyentes para el modelo mod.oxt4 con la transformación Box-Cox sobre la variable respuesta oxt.post.....	XIII
Figura V.D: comportamiento de los residuos del modelo mod.co.sngr sin aplicar ninguna transformación en la variable respuesta y en las covariables. Análisis gráfico de la linealidad, homocedasticidad, normalidad y valores outliers	XVIII
Figura VI.D: comportamiento de los residuos del modelo mod.co.sngr2 transformando logarítmicamente la variable respuesta y las covariables numéricas. Análisis gráfico de la linealidad, homocedasticidad, normalidad y valores outliers	XIX
Figura VII.D: comportamiento de los residuos del modelo mod.co.sngr4 aplicando la transformación Box-Cox sobre la variable respuesta. Análisis gráfico de la linealidad, homocedasticidad, normalidad y valores outliers	XX

Lista de tablas

Tabla I.A: cita y fecha de contacto de los artículos seleccionados para utilizar el conjunto de datos en el presente proyecto	IV
Tabla II.B: resultado del primer modelo planteado (mod.oxt) para predecir el nivel de oxitocina tras aplicar un estímulo.....	VII
Tabla III.B: sumario del resultado obtenido en el modelo mod.oxt3, donde la covariable edad no ha sido significativa al 5%.....	X
Tabla IV.B: resultado del output obtenido tras el planteamiento del modelo mod.oxt4, donde se observa que la variable edad no es significativa.....	XII
Tabla V.F: primeras seis observaciones de la transformación de la variable categórica etnia al definir el modelo de regresión en el software estadístico R obtenidas mediante la función model.matrix.....	XXIII
Tabla VI.F: resultado del output obtenido tras añadir la covariable etnia en el modelo que mide el nivel de la oxitocina tras aplicar un estímulo en el participante. Se observa que la variable predictora etnia no es significativa	XXIII
Tabla VII.F: resultado del output obtenido tras añadir la covariable etnia en el modelo que mide el nivel del cortisol tras aplicar un estímulo en el participante utilizando el conjunto de datos de la saliva. Se observa que la variable predictora etnia no es significativa.....	XXIV
Tabla VIII.F: resultado del output obtenido tras añadir la covariable etnia en el modelo que mide el nivel del cortisol tras aplicar un estímulo en el participante utilizando el conjunto de datos de la sangre. Se observa que la variable predictora etnia no es significativa.....	XXIV

Lista de ecuaciones

Eq. I.B: planteamiento del primer modelo (mod.oxt) para predecir el nivel de oxitocina tras aplicar un estímulo.....	VII
Eq. II.B: planteamiento del modelo mod.oxt3 para predecir el nivel de oxitocina tras aplicar un estímulo, transformando logarítmicamente la variable respuesta oxt.post	IX
Eq. III.B: ecuación para la transformación de la variable respuesta oxt.post para valores de lambda diferentes a cero	XI
Eq. IV.B: planteamiento del modelo mod.oxt4 con la transfromación de Box-Cox aplicada sobre la variable respuesta oxt.post	XII
Eq. V.C: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos completa) sin aplicar ninguna transformación en la variable respuesta (co.post) ni en las covariables	XV
Eq. VI.C: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos completa) aplicando la transformación logaritmica en las variables numéricas, tanto variable respuesta y en las covariables	XV
Eq. VII.C: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos completa) aplicando la transformación Box-Cox sobre la variable respuesta co.post	XVI
Eq. VIII.D: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos de las mediciones en sangre) sin aplicar ninguna transformación en la variable respuesta (co.post) ni en las covariables seleccionadas.....	XVII
Eq. IX.D: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos de las mediciones en sangre) transformando logarítmicamente la varible respuesta y las covariables numéricas seleccionadas	XVIII
Eq. X.D: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos de las mediciones en sangre) aplicando la transfromación Box-Cox sobre la variable respuesta co.post.....	XIX

Eq. XI.E: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos de las mediciones en saliva) sin aplicar ninguna transformación en la variable respuesta ni en las covariables	XXI
Eq. XII.E: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos de las mediciones en saliva) transformando logarítmicamente la variable respuesta co.post.....	XXI
Eq. XIII.E: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos de las mediciones en saliva) aplicando la transformación Box-Cox sobre la variable respuesta co.post.....	XXII

Anexo A

Generación base de datos

Tal y como se ha mencionado en el apartado 2.1 Generación de la base de datos, se contactó a 29 autores de artículos seleccionados con el objetivo de utilizar el conjunto de datos de su estudio para los análisis llevados a cabo en el presente proyecto. En la Tabla I.A que se muestra a continuación, se pueden observar los artículos seleccionados y la fecha en la que se contactó al autor/a correspondiente de los ensayos mediante correo electrónico.

Tabla I.A: cita y fecha de contacto de los artículos seleccionados para utilizar el conjunto de datos en el presente proyecto

Fecha de contacto (MM/DD/AAAA)	Cita APA artículos
08.26.2020	Tas, C., Brown, E. C., Eskikurt, G., Irmak, S., Aydin, O., Esen-Danaci, A., & Brüne, M. (2018). Cortisol response to stress in schizophrenia: associations with oxytocin, social support and social functioning. <i>Psychiatry research</i> , 270, 1047-1052. – <i>Respuesta 08.27.2020, derecho a utilizar los datos</i>
09.08.2020	Heinrichs, M., Baumgartner, T., Kirschbaum, C., & Ehlert, U. (2003). Social support and oxytocin interact to suppress cortisol and subjective responses to psychosocial stress. <i>Biological psychiatry</i> , 54(12), 1389-1398.
09.08.2020	Ditzen, B., Schaer, M., Gabriel, B., Bodenmann, G., Ehlert, U., & Heinrichs, M. (2009). Intranasal oxytocin increases positive communication and reduces cortisol levels during couple conflict. <i>Biological psychiatry</i> , 65(9), 728-731.
09.08.2020	Bhandari, R., Bakermans-Kranenburg, M. J., van der Veen, R., Parsons, C. E., Young, K. S., Grewen, K. M., ... & van IJzendoorn, M. H. (2014). Salivary oxytocin mediates the association between emotional maltreatment and responses to emotional infant faces. <i>Physiology & Behavior</i> , 131, 123-128.
09.14.2020	Atkinson, L., Gonzalez, A., Kashy, D. A., Santo Basile, V., Masellis, M., Pereira, J., ... & Levitan, R. (2013). Maternal sensitivity and infant and mother adrenocortical function across challenges. <i>Psychoneuroendocrinology</i> , 38(12), 2943-2951.
09.23.2020	Khoury, J. E., Gonzalez, A., Levitan, R., Masellis, M., Basile, V., & Atkinson, L. (2016). Maternal self-reported depressive symptoms and maternal cortisol levels interact to predict infant cortisol levels. <i>Infant Mental Health Journal</i> , 37(2), 125-139.
09.14.2020	Pierrehumbert, B., Torrisi, R., Laufer, D., Halfon, O., Ansermet, F., & Popovic, M. B. (2010). Oxytocin response to an experimental psychosocial challenge in adults exposed to traumatic experiences during childhood or adolescence. <i>Neuroscience</i> , 166(1), 168-177.
09.15.2020	Cardoso, C., Ellenbogen, M. A., Orlando, M. A., Bacon, S. L., & Joober, R. (2013). Intranasal oxytocin attenuates the cortisol response to physical stress: a dose-response study. <i>Psychoneuroendocrinology</i> , 38(3), 399-407.
09.15.2020	Alley, J., Diamond, L. M., Lipschitz, D. L., & Grewen, K. (2019). Associations between oxytocin and cortisol reactivity and recovery in response to psychological stress and sexual arousal. <i>Psychoneuroendocrinology</i> , 106, 47-56. – <i>Respuesta 09.16.2020, dicen que debo escribir a otros coautores en el artículo, que no respondieron.</i>

09.15.2020	Quirin, M., Kuhl, J., & Düsing, R. (2011). Oxytocin buffers cortisol responses to stress in individuals with impaired emotion regulation abilities. <i>Psychoneuroendocrinology</i> , 36(6), 898-904. – <i>Respuesta 09.17.2020, no envían el conjunto de datos</i>
09.15.2020	Luecken, L. J., MacKinnon, D. P., Jewell, S. L., Crnic, K. A., & Gonzales, N. A. (2015). Effects of prenatal factors and temperament on infant cortisol regulation in low-income Mexican American families. <i>Developmental psychobiology</i> , 57(8), 961-973. – <i>Respuesta 09.21.2020, no envían el conjunto de datos, solo datos agregados de su estudio</i>
09.15.2020	Frijling, J. L., van Zuiden, M., Nawijn, L., Koch, S. B. J., Neumann, I. D., Veltman, D. J., & Olff, M. (2015). Salivary oxytocin and vasopressin levels in police officers with and without post-traumatic stress disorder. <i>Journal of neuroendocrinology</i> , 27(10), 743-751.
09.15.2020	Grewen, K. M., Light, K. C., Mechlin, B., & Girdler, S. S. (2008). Ethnicity is associated with alterations in oxytocin relationships to pain sensitivity in women. <i>Ethnicity and Health</i> , 13(3), 219-241.
09.17.2020	Elmadih, A., Wan, M. W., Numan, M., Elliott, R., Downey, D., & Abel, K. M. (2014). Does oxytocin modulate variation in maternal caregiving in healthy new mothers?. <i>Brain research</i> , 1580, 143-150.
09.17.2020	Cong, X., Ludington-Hoe, S. M., Hussain, N., Cusson, R. M., Walsh, S., Vazquez, V., ... & Vittner, D. (2015). Parental oxytocin responses during skin-to-skin contact in pre-term infants. <i>Early Human Development</i> , 91(7), 401-406.
09.17.2020	Vittner, D., McGrath, J., Robinson, J., Lawhon, G., Cusson, R., Eisenfeld, L., ... & Cong, X. (2018). Increase in oxytocin from skin-to-skin contact enhances development of parent–infant relationship. <i>Biological research for nursing</i> , 20(1), 54-62.
09.17.2020	Samuel, S., Hayton, B., Gold, I., Feeley, N., Carter, C. S., & Zelkowitz, P. (2015). Maternal mental health moderates the relationship between oxytocin and interactive behavior. <i>Infant mental health journal</i> , 36(4), 415-426.
09.17.2020	Kory Floyd, Alan C. Mikkelson, Melissa A. Tafoya, Lisa Farinelli, Angela G. La Valley, Jeff Judd, Mark T. Haynes, Kristin L. Davis & Jason Wilson (2007) Human Affection Exchange: XIII. Affectionate Communication Accelerates Neuroendocrine Stress Recovery, <i>Health Communication</i> , 22:2, 123-132 – <i>Respuesta 09.21.2020, no envían el conjunto de datos</i>
09.18.2020	Suzuki, S., Fujisawa, T. X., Sakakibara, N., Fujioka, T., Takiguchi, S., & Tomoda, A. (2020). Development of Social Attention and oxytocin Levels in Maltreated children. <i>Scientific Reports</i> , 10(1), 1-10.
09.18.2020	Fujisawa, T. X., Tanaka, S., Saito, D. N., Kosaka, H., & Tomoda, A. (2014). Visual attention for social information and salivary oxytocin levels in preschool children with autism spectrum disorders: an eye-tracking study. <i>Frontiers in neuroscience</i> , 8, 295.
09.18.2020	Bellosta-Batalla, M., Blanco-Gandía, M. D. C., Rodríguez-Arias, M., Cebolla, A., Pérez-Blasco, J., & Moya-Albiol, L. (2020). Brief mindfulness session improves mood and increases salivary oxytocin in psychology students. <i>Stress and Health</i> . – <i>Respuesta 09.18.2020, no envían el conjunto de datos, solo datos agregados del estudio</i> .
09.24.2020	Li, Y., Hassett, A. L., & Seng, J. S. (2019). Exploring the mutual regulation between oxytocin and cortisol as a marker of resilience. <i>Archives of psychiatric nursing</i> , 33(2), 164-173.

10.05.2020	Kubzansky, L. D., Mendes, W. B., Appleton, A. A., Block, J., & Adler, G. K. (2012). A heartfelt response: oxytocin effects on response to social stress in men and women. <i>Biological psychology</i> , 90(1), 1-9.
10.05.2020	Naber, F., van IJzendoorn, M. H., Deschamps, P., van Engeland, H., & Bakermans-Kranenburg, M. J. (2010). Intranasal oxytocin increases fathers' observed responsiveness during play with their children: a double-blind within-subject experiment. <i>Psychoneuroendocrinology</i> , 35(10), 1583-1586.
10.18.2020	Bischoff, M., Howland, V., Klinger-König, J., Tomczyk, S., Schmidt, S., Zygmunt, M., ... & Günther, S. (2019). Save the children by treating their mothers (PriVileG-M-study)-study protocol: a sequentially randomized controlled trial of individualized psychotherapy and telemedicine to reduce mental stress in pregnant women and young mothers and to improve Child's health. <i>BMC psychiatry</i> , 19(1), 1-13.
10.18.2020	Tanaka, S., Komagome, A., Iguchi-Sherry, A., Nagasaka, A., Yuhi, T., Higashida, H., ... & Tsuji, T. (2020). Participatory Art Activities Increase Salivary Oxytocin Secretion of ASD Children. <i>Brain Sciences</i> , 10(10), 680.
10.18.2020	Hood, C. O., Tomko, R. L., Baker, N. L., Tuck, B. M., Flanagan, J. C., Carpenter, M. J., ... & McClure, E. A. (2020). Examining sex, adverse childhood experiences, and oxytocin on neuroendocrine reactivity in smokers. <i>Psychoneuroendocrinology</i> , 104752.

Anexo B

Modelo oxitocina con el conjunto de datos completo

En el presente Anexo B se describen los diferentes modelos planteados para la oxitocina. Se trata del modelo *mod.oxt* (sin ninguna transformación en las variables), *mod.oxt3* (donde únicamente se ha transformado logarítmicamente la variable respuesta) y *mod.oxt4*, donde se ha aplicado la transformación Box-Cox sobre la variable respuesta.

Modelo I

El modelo I se describe con la variable dependiente *oxt.post* y las cuatro variables predictoras (tres de ellas numéricas y una categórica). El modelo *mod.oxt*, es el primero planteado para la oxitocina, pero los resultados obtenidos no han sido adecuados para utilizarlo como predictor del nivel de oxitocina. El modelo se plantea de la siguiente manera:

$$Y = B_0 + B_1 (X_{age}) + B_2 (X_{stimulus.type}) + B_3 (X_{oxt.pre}) + B_4 (X_{hr.bas}) + \epsilon$$

Eq. I.B: planteamiento del primer modelo (mod.oxt) para predecir el nivel de oxitocina tras aplicar un estímulo

Tras su definición en R, el resultado obtenido del sumario del modelo se muestra en la Tabla II.B que se muestra a continuación:

Tabla II.B: resultado del primer modelo planteado (mod.oxt) para predecir el nivel de oxitocina tras aplicar un estímulo

Predictores	Coeficiente B	Std.Err	t	Sig
constante	-0.73982	2.30185	-0.321	0.74953
edad	-0.14661	0.07217	-2.032	0.0487*
stimulus.type	-1.39790	0.51116	-2.735	0.009179**
oxt.pre	1.03387	0.08080	12.796	6.63e-16***
hr.bas	0.09195	0.02563	3.587	0.000882 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
F	43.64			
R ²	0.7912			
p-valor	2.952e-14			

Del resumen obtenido mediante la función *summary* del modelo planteado se observa que todas las variables explicativas son significativas al 5%, aunque la variable *age* se encuentre en el límite para considerarse significativa, con un p-valor=0.049. El valor del *R²* ajustado es de 0.7912, considerado elevado. Debido al p-valor ajustado, es adecuado analizar si eliminar la variable *age* mejoraría el modelo, aunque esto hay que confirmarlo mediante un test. Para ver si efectivamente debería eliminarse la variable edad del análisis, se lleva a cabo Akaike, que mide el ajuste del modelo utilizando la función *stepAIC* sobre el mismo.

El análisis de Akaike ha determinado que la variable predictora *age*, aunque sea la que menos modificaría los resultados del modelo en caso de que fuera eliminada, sí que se considera relevante para el modelo, y por lo tanto, se mantiene. Sin embargo, es necesario analizar si los residuos del modelo cumplen con las condiciones necesarias:

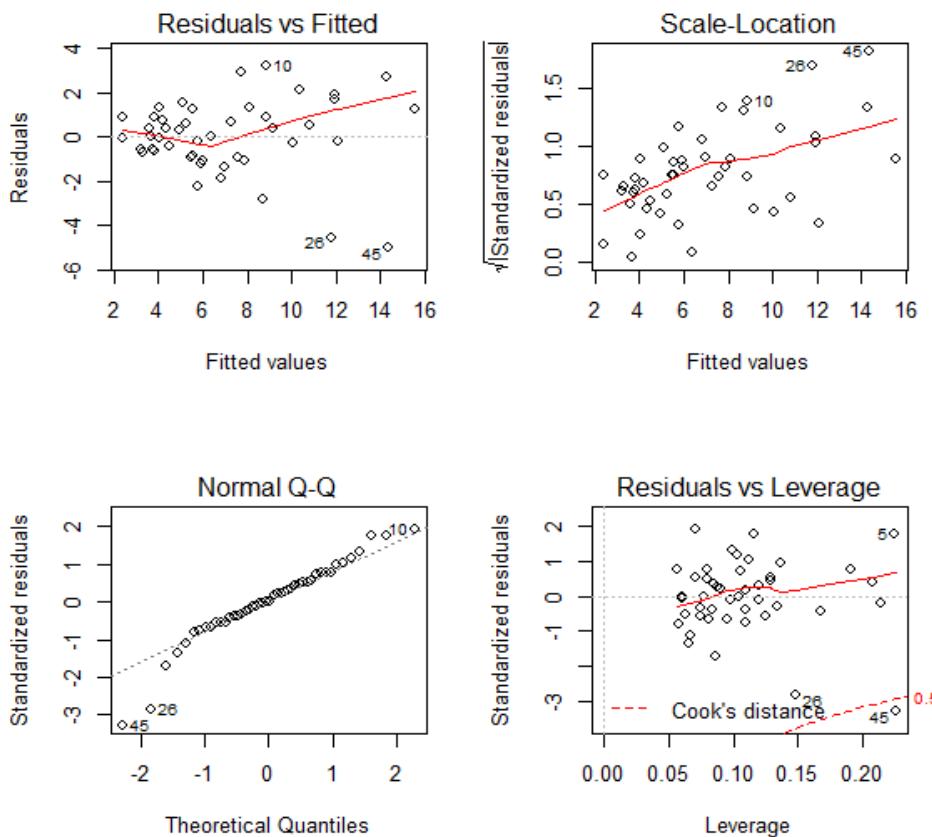


Figura I.B: residuos del primer modelo planteado mod.oxt para predecir el nivel de oxitocina tras aplicar un estímulo.
Gráfico de linealidad (arriba izq.), homocedasticidad (arriba dcha.), normalidad (abajo izq.) y puntos outliers o influyentes (abajo dcha.)

Tal y como se ha explicado para el modelo *mod.oxt2* en el documento, cada uno de los gráficos analiza diferentes aspectos en relación a los residuos del modelo. Se trata de la linealidad, normalidad, homocedasticidad/heterocedasticidad y valores influyentes (*outliers*), tal y como se describen en los siguientes puntos.

- Linealidad: analizado en el gráfico *Residuals vs Fitted*, que muestra si el modelo es una combinación lineal de las variables predictoras. En el modelo *mod.oxt*, se observa que este principio no se cumple, ya que la línea roja no se está sobreponiendo a la línea horizontal central.
- Normalidad: analizado en el gráfico *Normal Q-Q*, que muestra si los residuos están distribuidos de forma normal. En este caso, observamos que las colas no están del todo alineadas con la línea central, por lo tanto a simple vista no es posible saber si el principio de normalidad se cumple o no, aunque se observa que la mayoría de puntos centrales sí que están sobre la línea.
- Homocedasticidad: analizado en el gráfico *Scale Location*, que muestra si la varianza de los residuos está distribuida de forma constante para las variables predictoras. En este caso se observa que la línea roja no es horizontal (por lo que puede ser que los residuos vayan cambiando para los valores predichos) y la distribución alrededor de la línea roja cuando los valores en el eje x (*fitted values*) aumentan parece que varían. El término contrario a la homocedasticidad es la heterocedasticidad, que sería el supuesto de que la varianza de los residuos no es constante, como parece ser el caso para el modelo *mod.oxt*.

- Detectar valores influyentes (*outliers*) del modelo: mediante el gráfico *Residuals vs Leverage*. Los valores que se muestran separados del resto mediante la línea discontinua, son valores influyentes, que de eliminarlos, el comportamiento del modelo cambiaría (normalmente mejorándolo). En este caso, se observa que existe una observación (la 45) separada por la distancia de Cook.

Es necesario verificar estas suposiciones mediante diferentes tests sobre los residuos del modelo *mod.oxt*.

Normalidad de los residuos:

Lo primero que se deberá hacer será verificar mediante un test de normalidad si los residuos del modelo *mod.oxt* siguen o no una distribución normal, ya que gráficamente (en el gráfico Q-Q), se ha observado que las colas difieren de lo que se consideraría una distribución normal. Para comprobar la normalidad, se aplica la función *Shapiro.test* del paquete *MASS* que hace referencia al test Shapiro-Wilk. Este test, asume en su hipótesis nula que los residuos siguen una distribución normal.

En el test se obtiene un p-valor=0.05, justo en el límite del nivel de significancia establecido en el estudio, aunque no es evidencia suficiente para rechazar la hipótesis nula y por lo tanto, se asume la normalidad de los residuos.

Homocedasticidad/heterocedasticidad:

Se analiza la homocedasticidad/heterocedasticidad del modelo *mod.oxt* utilizando el test *Non-Constant Variance Score Test (ncVs)* y el test Breusch-Pagan, tal y como se ha explicado en el apartado 2.3.5.2 del documento. Ambos tests asumen en su hipótesis nula que la varianza de los residuos es constante y en la hipótesis alternativa que la varianza cambia según los valores ajustados o la combinación lineal de variables predictoras. En los resultados de ambos tests se obtiene un p-valor inferior que el nivel de significancia al 5% (p=3.3805e-06 y p=0.003258 respectivamente), por lo tanto se rechaza la hipótesis nula y no se podría determinar que la varianza de los residuos del modelo es constante ya que se asume la existencia de la heterocedasticidad.

Como no se ha cumplido la suposición de homocedasticidad para el modelo *mod.oxt*, necesario para un modelo lineal, este modelo se ha rechazado y se han planteado diferentes transformaciones de las variables, tal y como se explica en las siguientes subsecciones. Además, también se intentará que la condición de linealidad observada en los gráficos de los residuos mejore.

Modelo II

El siguiente modelo que se plantea es el modelo *mod.oxt3*, donde únicamente se modifica la variable respuesta (*oxt.post*), transformándola en una variable logarítmica. El modelo se denomina *mod.oxt3* y su planteamiento se muestra a continuación:

$$\log(Y) = B_0 + B_1 (X_{age}) + B_2 (X_{stimulus.type}) + B_3 (X_{oxt.pre}) + B_4 (X_{hr.bas}) + \epsilon$$

Eq. II.B: planteamiento del modelo mod.oxt3 para predecir el nivel de oxitocina tras aplicar un estímulo, transformando logarítmicamente la variable respuesta oxt.post

Tras aplicarlo en R, el resultado obtenido del sumario del modelo, se muestra en la Tabla III.B:

Tabla III.B: sumario del resultado obtenido en el modelo mod.oxt3, donde la covariable edad no ha sido significativa al 5%

Predictores	Coeficiente B	Std.Err	t	Sig
Constante	0.554134	0.323648	1.712	0.094424
Edad	-0.017533	0.010147	-1.728	0.091526
Stimulus.type2	-0.177136	0.071872	-2.465	0.017996 *
Oxt.pre	0.138581	0.011361	12.198	3.16e-15 ***
Hr.bas	0.014600	0.003604	4.051	0.000221 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
F	40.57			
R ²	0.7786			
p-valor	2.679e-14			

Tal y como se muestra en la Tabla III.B para analizar el sumario del modelo, se observa que la variable *age* no es significativa al 5% (p valor= 0.091), por lo que podría considerarse que se debería eliminar del modelo. Sin embargo, al realizar Akaike, aunque sí que sea la variable que menos influencia tiene sobre la respuesta, éste no aconseja su eliminación (además tiene un p-valor cercano a 0.05), por lo que se mantiene en el modelo. Además, el valor del *R*² ajustado es más bajo que para el modelo *mod.oxt* descrito arriba y el modelo *mod.oxt2* descrito en el apartado 2.3.5 del documento. Aunque el valor de *R*² ajustado sea más bajo, también se analiza el comportamiento de los residuos para los diferentes supuestos del modelo, tal y como se observa en la Figura II.B:

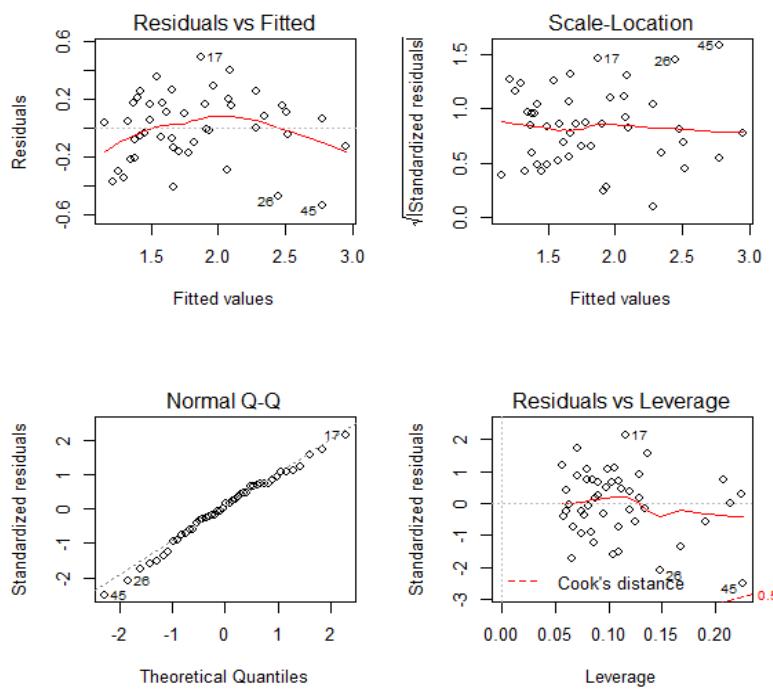


Figura II.B: análisis de los residuos en términos de linealidad, homocedasticidad, normalidad y valores influyentes para el modelo mod.oxt3

A simple vista, se observa que la linealidad no se cumple ya que la línea roja no es horizontal y no está sobreposta en la línea central. Respecto a la normalidad, una vez más las colas parecen que difieren de la línea central. Existen puntos *outliers* (aunque ninguno distanciado por Cook), y finalmente, en el gráfico de *scale-location* no es posible a simple vista determinar si se cumple o no la homocedasticidad, aunque una vez más se observan que para los valores más altos los residuos están más dispersos. Estos supuestos se analizan aplicando los tests descritos en el apartado 2.3.5 del documento para el modelo *mod.oxt2*.

Normalidad de los residuos:

La normalidad de los residuos se ha analizado aplicando el test de Shapiro-Wilk sobre ellos. Se ha obtenido un p-valor = 0.855, por lo tanto no hay evidencia suficiente para rechazar la hipótesis nula cuya definición se basa en la normalidad de los residuos.

Homocedasticidad/heterocedasticidad:

Se analiza la homocedasticidad/heterocedasticidad del modelo *mod.oxt3* utilizando una vez más los tests *Non-Constant Variance Score Test (ncVs)* y Breusch-Pagan, tal y como se ha explicado en el documento previo. De los resultados de ambos se obtiene que no existe evidencia suficiente para rechazar la hipótesis nula de los dos tests, por lo tanto se puede aceptar que la varianza es constante para los residuos del modelo *mod.oxt3* (p-valor = 0.387 y p-valor= 0.6 respectivamente).

Finalmente, aunque la suposición de normalidad, homocedasticidad, no multicolinealidad y no autocorrelación se acepten para los residuos de este modelo, el gráfico de linealidad mostrado (*Residuals vs Fitted*) de la Figura II.B no muestra un comportamiento ideal. Además, al obtener un valor del R^2 ajustado inferior que para los demás modelos, ésta transformación ha sido rechazada para predecir el nivel de oxitocina tras aplicar un estímulo sobre un paciente.

Modelo III

La siguiente transformación que se muestra es la transformación Box-Cox aplicada sobre la variable respuesta *oxt.post*. La transformación de Box-Cox se suele aplicar para que los residuos del modelo se asemejen a una distribución normal y también para mejorar la linealidad de los residuos. Se ha observado que los residuos de los modelos sí que siguen hasta ahora una distribución normal, y en el presente subapartado se analiza si la transformación Box-Cox sobre la variable respuesta mejora el modelo en relación a la linealidad.

Antes de aplicar la transformación, es necesario conocer cómo se realiza la transformación de la variable respuesta *Y* cuando λ es diferente a cero y la variable respuesta es positiva. La transformación se muestra a continuación:

$$y(\lambda) = \frac{y^\lambda - 1}{\lambda}$$

*Eq. III.B: ecuación para la transformación de la variable respuesta *oxt.post* para valores de lambda diferentes a cero*

Cuando λ es cero, la transformación que se lleva a cabo es la misma que se ha mostrado en el subapartado anterior “Modelo II” de este mismo Anexo.

Antes de aplicar la transformación, se debe calcular el valor máximo de *lambda* sobre el modelo *mod.oxt* (sin transformar). Gráficamente, se puede obtener una estimación del valor de λ para el modelo *mod.oxt*, tal y como se muestra en la Figura III.B:

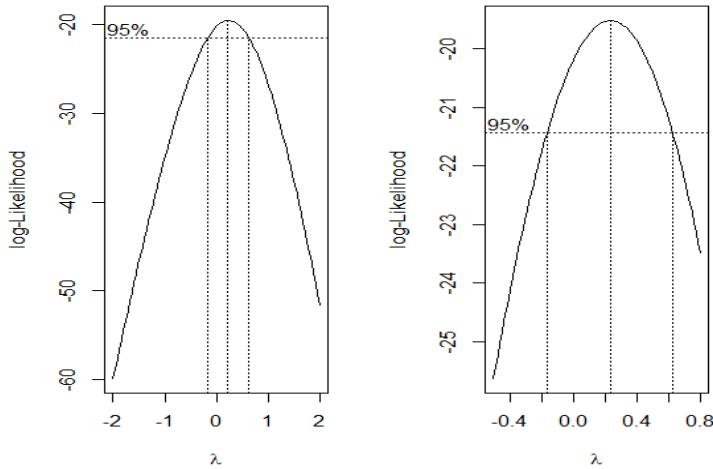


Figura III.B: estimación del valor de *lambda* para el modelo que predice el nivel de oxitocina post aplicación de un estímulo sin ninguna transformación (mod.oxt). En el gráfico de la izquierda, se muestra el intervalo de confianza para el valor de *lambda*, y en la derecha se observa que el valor es cercano a 0.20

En el gráfico de la izquierda se observa que el valor de *lambda* máximo se encuentra entre los valores 0 y 1 en un intervalo de confianza del 95% y en el gráfico de la derecha, se observa que el valor es cercano a 0.25 aproximadamente (también con un intervalo de confianza del 95%). Aplicando la función *which.max* se conoce que el valor máximo de *lambda* (λ) es 0.222 para el modelo mod.oxt. Estos valores se deben sustituir en la fórmula de la transformación Box-Cox mostrada previamente para la variable respuesta. El modelo planteado se denomina mod.oxt4, con el valor de $\lambda = 0.222$. La formula es la siguiente:

$$\frac{Y^\lambda - 1}{\lambda} = B_0 + B_1 (X_{age}) + B_2 (X_{stimulus.type}) + B_3 (X_{oxt.pre}) + B_4 (X_{hr.bas}) + \epsilon$$

Eq. IV.B: planteamiento del modelo mod.oxt4 con la transformación de Box-Cox aplicada sobre la variable respuesta ox.t.post

El output del resumen obtenido en R tras aplicar la formula se muestra en la Tabla IV.B:

Tabla IV.B: resultado del output obtenido tras el planteamiento del modelo mod.oxt4, donde se observa que la variable edad no es significativa

Predictores	Coeficiente B	Std.Err	t	Sig
Constante	0.412335	0.478020	0.863	0.393380
Edad	-0.027768	0.014987	-1.853	0.071117
Stimulus.type2	-0.276589	0.106152	-2.606	0.012724 *
Oxt.pre	0.213207	0.016779	12.707	8.35e-16 ***
Hr.bas	0.021765	0.005323	4.089	0.000197 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1				
F	43.79			
R ²	0.7918			
p-valor	2.782e-14			

En la Tabla IV.B se observa que la variable predictora *age* no es significativa aunque el p-valor sea 0.07 (muy cercano el 5% del nivel de significancia establecido durante todo el estudio). Se aplica la función *stepAIC* para analizar si se debe mantener o no la variable predictora *age*, y en base a los resultados obtenidos mediante Akaike, la variable predictora *age* debe mantenerse en el modelo, aunque no sea significativa al 5%.

Una vez más, es necesario comprobar gráficamente y posteriormente utilizando los diferentes tests cómo se comportan los residuos en este modelo. Los gráficos se muestran a continuación en la Figura IV.B:

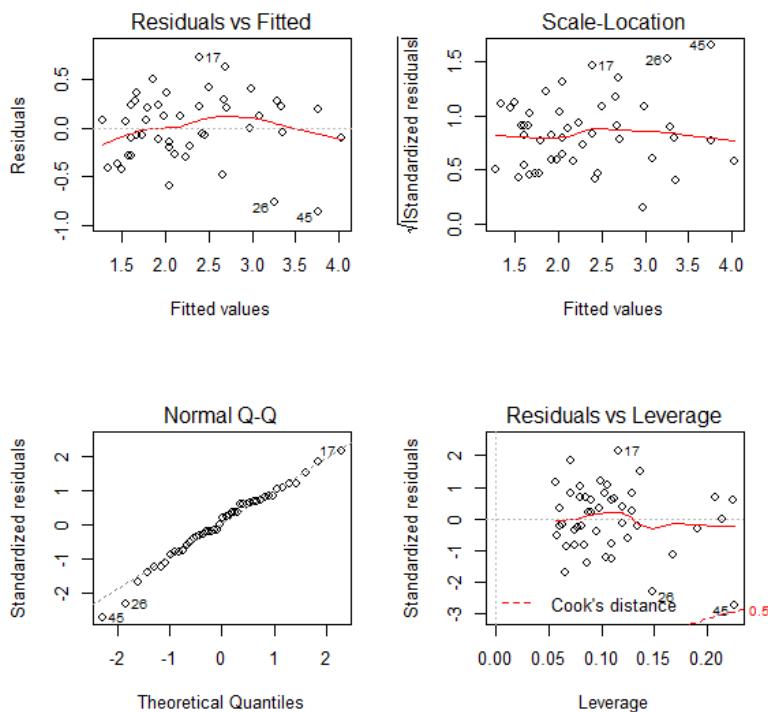


Figura IV.B: análisis de los residuos en términos de linealidad, homocedasticidad, normalidad y valores influyentes para el modelo mod.oxt4 con la transformación Box-Cox sobre la variable respuesta oxt.post

Gráficamente parece que la normalidad sigue teniendo un comportamiento bastante parecido que en los casos anteriores, ya que se observan residuos más alejados en la zona de las colas (gráfico QQ). En relación a la homocedasticidad (gráfico Scale-Location), parece que existe mayor dispersión respecto a la línea roja para los valores más altos, pero habrá que analizarlo mediante un test, para aceptar o rechazar finalmente la homocedasticidad de los residuos. En relación a la linealidad, parece que ésta, a simple vista se cumple y que se obtienen mejores resultados que al menos en los anteriores modelos mostrados en el presente Anexo. En relación a los puntos *outliers*, se sigue observando que hay algunos pero ninguno de ellos está fuera de la distancia de Cook. A continuación se llevan a cabo los tests para analizar las suposiciones.

Normalidad de los residuos:

Utilizando el test de Shapiro-Wilk, se lleva a cabo el análisis de la normalidad para el modelo mod.oxt4, y dado que la hipótesis nula acepta la normalidad de los residuos, y como se ha obtenido un p-valor de 0.8037, no hay evidencia suficiente para rechazar la hipótesis nula, por lo tanto se asume la normalidad de los residuos.

Homocedasticidad/heterocedasticidad:

Es posible analizar la existencia de heterocedasticidad tal y como se ha hecho previamente utilizando el test *Non-Constant Variance Score Test (ncVs)* o el Breusch-Pagan Test, aplicando la función *ncvTest* o *bptest* respectivamente sobre el modelo. Ambos tests, asumen en su hipótesis nula que la varianza de los residuos es constante. En este caso, no hay evidencia suficiente (ya que se obtiene un valor de mayor que 0.05 para ambos tests) para rechazar la hipótesis nula, y

por ello se acepta que la varianza de los residuos es constante, y se asume que los residuos son homocedásticos.

Autocorrelación:

Para analizar la autocorrelación entre las variables, en este caso se ha aplicado también el test de *Durbin-Watson* tal y como se ha hecho para las transformaciones anteriores. El test se aplica mediante la función *durbinWatsonTest* sobre el modelo *mod.oxt4*, y en el *output* obtenido se observa que el p-valor=0.524, y que por lo tanto se asume que las variables son independientes ya que no hay evidencia suficiente para rechazar la hipótesis nula.

Multicolinealidad:

En este caso también se analiza la multicolinealidad mediante el test de *Farrar - Glauber* para observar si existe multicolinealidad entre las variables predictoras del *mod.oxt4*, y como todos los valores del *Klein* en el resultado se igualan a cero, se asume que no se ha detectado colinealidad. Además, mediante la función *vif* - *Variance inflation factor*, que cuantifica la correlación entre las variables predictoras de un modelo, se ha observado que las cuatro variables predictoras tienen valores pequeños, cercanos a uno (mínimo 1.01 y máximo 1.19), por lo tanto no parece que exista colinealidad entre éstas variables.

Anexo C

Modelo cortisol con el conjunto de datos completo

En el presente Anexo C se describen los diferentes modelos planteados para el biomarcador cortisol utilizando la base de datos generada. Se describen los modelos *mod.co.p1* (sin ninguna transformación en la variable respuesta ni en las variables predictoras), *mod.co.p3* (transformando logarítmicamente la variable respuesta), y *mod.co.p4* (transformación BoxCox sobre la variable respuesta).

Modelo I

El modelo *mod.co.p1* se ha definido con la variable respuesta *co.post* y en un principio con las variables predictoras *age*, *gender*, *stimulus.type*, *co.pre*, *co.reac* y *hr.post*, tal y como se muestra a continuación:

$$Y = B_0 + B_1 (X_{age}) + B_2 (X_{gender}) + B_3 (X_{stimulus.type}) + B_4 (X_{co.pre}) \\ + B_5 (X_{co.reac}) + B_6 (X_{hr.post}) + \epsilon$$

*Eq. V.C: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos completa) sin aplicar ninguna transformación en la variable respuesta (*co.post*) ni en las covariables*

Sin embargo, como se ha explicado en el documento, la variable *hr.post* únicamente se ha medido en uno de los artículos, y por lo tanto tiene un gran porcentaje de valores faltantes. Por lo tanto, se ha eliminado la variable de *hr.post* en el planteamiento de los modelos. Tras el planteamiento con los valores predictores con las variables *age*, *gender*, *stimulus.type*, *co.pre* y *co.reac*, se ha observado que las variables significativas son *stimulus.type*, *co.pre* y *co.reac*, y el R^2 es 0.967, con un valor muy significativo al 5%. Respecto a los residuos del modelo, gráficamente se observa linealidad, pero respecto a la varianza de los residuos, no se observa que sea constante, y además, al aplicar los test, los p-valores obtenidos son menores que 0.05. Al comparar el modelo con los otros tres planteados, se ha observado un valor AIC y BIC más alto que con los demás.

Modelo II

El modelo *mod.co.p2* se ha definido con la variable respuesta *co.post* y en un principio con las variables predictoras *age*, *gender*, *stimulus.type*, *co.pre*, *co.reac* y *hr.post*, transformando logarítmicamente las variables numéricas.

$$\log(Y) = B_0 + B_1 \log(X_{age}) + B_2 \log(X_{gender}) + B_3 \log(X_{stimulus.type}) + B_4 \log(X_{co.pre}) \\ + B_5 \log(X_{co.reac}) + B_6 \log(X_{hr.post}) + \epsilon$$

Eq. VI.C: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos completa) aplicando la transformación logarítmica en las variables numéricas, tanto variable respuesta y en las covariables

Se ha tenido que eliminar la variable *hr.post* del modelo, debido a los valores faltantes que hay en los conjuntos de datos, tal y como se ha explicado en el documento. Tras ejecutar el modelo, únicamente han resultado significativas las variables *co.pre* y *co.post*, ambas logarítmicamente transformadas. Tras aplicar Akaike, efectivamente se ha confirmado que solo había que incluir las dos variables mencionadas, y el valor del R^2 obtenido ha sido de 0.9592. En el análisis de los residuos, se ha observado que a simple vista no parece que se cumpla la suposición de normalidad en los residuos, y así se ha confirmado mediante el test de Shapiro-Wilk, con un p-valor muy por debajo del nivel de significancia del 5%. Sin embargo, tanto gráficamente como

mediante los dos test que se han ido aplicando para el análisis de la homocedasticidad, sí que se ha observado que la varianza de los residuos es constante. Finalmente, comentar que también se observan valores *outliers* y que la linealidad no se cumple del todo.

Modelo III

El último modelo que se ha planteado con los datos del conjunto de datos del cortisol se denomina *mod.co.p4*, y en este caso se ha aplicado la transformación Box-Cox sobre la variable respuesta *co.post*. Del mismo que para el biomarcador *oxitocina*, primero se ha calculado el valor de *lambda* a partir del modelo sin ninguna transformación. Se ha obtenido un valor de *lambda* = 0.70, y éste se ha aplicado sobre la variable respuesta *co.post* mediante la función $y(\lambda) = \frac{y^\lambda - 1}{\lambda}$ sobre ella. El modelo planteado en un principio se describe en la siguiente función (eliminando la variable predictora *hr.post*):

$$\frac{Y^\lambda - 1}{\lambda} = B_0 + B_1 (X_{age}) + B_2 (X_{gender}) + B_3 (X_{co.pre}) + B_4 (X_{stimulus.type}) \\ + B_5 (X_{co.reac}) + \epsilon$$

Eq. VII.C: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos completa) aplicando la transformación Box-Cox sobre la variable respuesta co.post

En este modelo, las variables significativas han sido *co.pre*, *co.reac* y un nivel (igualado a uno) de la variable *stimulus.type*. También se ha aplicado Akaike, y pese a que *stimulus.type*=2 no fuera significativo, la variable se debe mantener en el modelo. El valor de R^2 obtenido es muy alto, igualado a 0.9719. Sin embargo, en relación a los residuos del modelo, no se cumple con la suposición de normalidad ni con la de homocedasticidad, por lo tanto el modelo se ha rechazado para el análisis del cortisol post estímulo utilizando la base de datos completa de este biomarcador.

Anexo D

Modelo cortisol con el conjunto de datos con mediciones en sangre

En el presente Anexo C se describen los diferentes modelos planteados para el biomarcador cortisol utilizando la base de datos de las mediciones realizadas a partir de las muestras de sangre. Se describen los modelos *mod.co.sngr* (sin ninguna transformación en la variable respuesta ni en las variables predictoras), *mod.co.sngr2* (transformando logarítmicamente todas las variables numéricas, respuesta y predictoras), y *mod.co.sngr4* (transformación BoxCox).

Modelo I

El modelo *mod.co.sngr* se ha definido con la variable respuesta *co.post* y las variables predictoras *co.pre*, *age*, *co.reac*, *med.dos* y *gender*, tal y como se muestra a continuación:

$$Y = B_0 + B_1 (X_{co.pre}) + B_2 (X_{age}) + B_3 (X_{co.reac}) + B_4 (X_{med.dos}) + B_5 (X_{gender}) + \epsilon$$

Eq. VIII.D: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos de las mediciones en sangre) sin aplicar ninguna transformación en la variable respuesta (co.post) ni en las covariables seleccionadas

Sin embargo, no todas las variables predictoras han resultado ser significativas, y tras aplicar Akaike para determinar qué variables generan una influencia sobre la variable respuesta, se ha determinado que únicamente deberían incluirse las variables *co.pre* y *co.reac*. Aunque el R^2 obtenido en el modelo sea muy elevado ($R^2 = 0.95$), el modelo no cumple con las suposiciones de la linearidad. Gráficamente (tal y como se muestra en la Figura V.D), se observa que los residuos del modelo no son homocedásticos (se ha generado una forma de *campana*) ni tampoco cumplen el supuesto de la linealidad. Además, al aplicar el test de Shapiro-Wilk para la normalidad, se ha observado que no se acepta la hipótesis nula de normalidad puesto que se obtiene un p-valor inferior al 5%. Lo mismo ocurre con la normalidad, ya que con ninguno de los dos test aplicados se obtiene un p-valor superior al 5%, por lo que tal y como se había intuido gráficamente, la varianza de los residuos es heterocedástica.

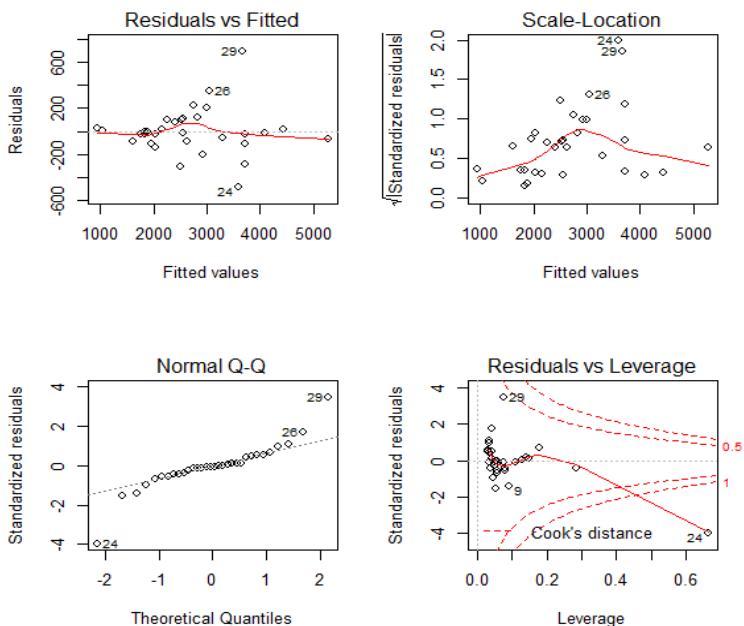


Figura V.D: comportamiento de los residuos del modelo *mod.co.sngr* sin aplicar ninguna transformación en la variable respuesta y en las covariables. Análisis gráfico de la linealidad, homocedasticidad, normalidad y valores outliers

Los resultados observados en los gráficos de la Figura V.D y los resultados de los test son suficientes para descartar el modelo *mod.co.sngr* para predecir el nivel de cortisol post estímulo utilizando las muestras de sangre.

Modelo II

El segundo modelo planteado se ha denominado *mod.co.sngr2* y en él se han transformado logarítmicamente todas las variables numéricas, tal y como se observa a continuación:

$$\begin{aligned} \log(Y) = & B_0 + B_1 \log(X_{co.pre}) + B_2 \log(X_{age}) + B_3 \log(X_{co.reac}) \\ & + B_4 \log(X_{med.dos}) + B_5 X_{gender} + \epsilon \end{aligned}$$

Eq. IX.D: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos de las mediciones en sangre) transformando logarítmicamente la variable respuesta y las covariables numéricas seleccionadas

Del mismo modo que para el modelo anterior (*mod.co.sngr1*), al aplicar Akaike sobre el modelo, únicamente se han mantenido las variables significativas al 5%, las cuales han sido las variables *co.pre* y *co.reac*, esta vez transformadas logarítmicamente. El modelo ha mantenido un valor del *R*² ajustado alto (con un valor de 0.76), pero tampoco se han cumplido los supuestos necesarios para aceptar finalmente el modelo. Al aplicar el test de normalidad sobre él, se ha obtenido un p-valor inferior al 5%, y en el caso de la homocedasticidad, el test *ncVs* no ha sido significativo (*p=0.02*) pero por el contrario, el test *Breusch-Pagan* sí. Gráficamente, el comportamiento de los residuos del modelo se observa a continuación:

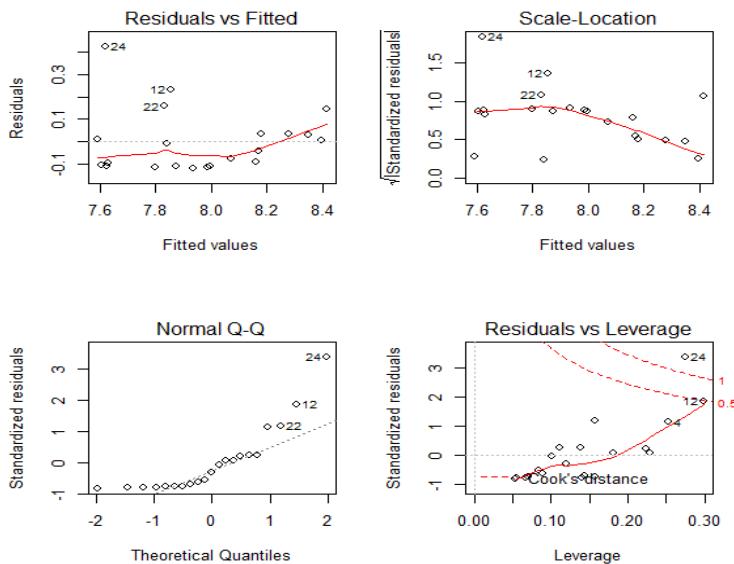


Figura VI.D: comportamiento de los residuos del modelo *mod.co.sngr2* transformando logarítmicamente la variable respuesta y las covariables numéricas. Análisis gráfico de la linealidad, homocedasticidad, normalidad y valores outliers

A parte de los resultados obtenido en los test, los resultados gráficos observados en la Figura VI.D son suficientes para descartar el modelo *mod.co.sngr2*, ya que tampoco se cumple la linealidad de los residuos, y en el caso de la homocedasticidad, gráficamente no hay evidencia suficiente para aceptarla, aunque en uno de los tests se haya obtenido un p-valor superior al 5%.

Modelo III

El último modelo que se ha planteado con los datos de las mediciones en sangre se denomina *mod.co.sngr4*, y en este caso se ha aplicado la transformación Box-Cox sobre la variable respuesta *co.post*. Del mismo modo que para el biomarcador *oxitocina*, primero se ha calculado el valor de *lambda* a partir del modelo sin ninguna transformación. Se ha obtenido un valor de *lambda* = 0.86, y éste se ha aplicado sobre la variable respuesta *co.post* utilizando la función $y(\lambda) = \frac{y^\lambda - 1}{\lambda}$ sobre ella. El modelo planteado en un principio se describe en la siguiente función:

$$\frac{Y^\lambda - 1}{\lambda} = B_0 + B_1 (X_{co.pre}) + B_2 (X_{age}) + B_3 (X_{co.reac}) + B_4 (X_{med.dos}) + B_5 (X_{gender}) + \epsilon$$

Eq. X.D: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos de las mediciones en sangre) aplicando la transformación Box-Cox sobre la variable respuesta co.post

En este caso, como para los modelos anteriores, también se ha aplicado la función de Akaike para determinar qué variables debían mantenerse según el efecto obtenido sobre la variable respuesta y la significancia en el modelo. Según el método Akaike, únicamente se han debido mantener las variables predictoras *co.pre*, *med.dos* y *co.reac*, aunque la variable *med.dos* no haya obtenido un p-valor significativo al 5%. Una vez más, el R^2 del modelo ha sido muy alto, con un valor de 0.96. Aunque el test de normalidad de Shapiro-Wilk haya aceptado la normalidad de los datos, los resultados en los test de homocedasticidad no han resultado significativos al 5%, y por lo tanto, existe evidencia suficiente para rechazar este modelo que predice la variable respuesta *co.post*. En la Figura VII.D se muestra el comportamiento de los residuos del modelo,

donde se observa en el gráfico *scale location* que la varianza de los residuos no es constante debido a la forma acampanada que se genera. Sin embargo, cabe destacar que la linealidad para los residuos de este modelo parece adecuada, y que los residuos están distribuidos de forma normal a pesar de los valores *outliers* observados en ambas colas. Sin embargo, tal y como se ha comentado, el modelo queda descartado para el análisis.

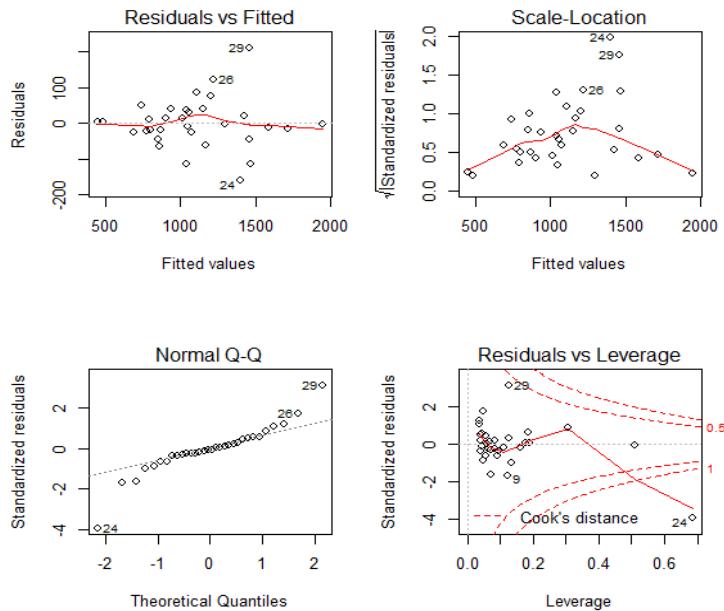


Figura VII.D: comportamiento de los residuos del modelo `mod.co.sngr4` aplicando la transformación Box-Cox sobre la variable respuesta. Análisis gráfico de la linealidad, homocedasticidad, normalidad y valores outliers

Anexo E

Modelo cortisol con el conjunto de datos con mediciones en saliva

En el presente Anexo E se describen los diferentes modelos planteados para el biomarcador cortisol utilizando la base de datos de las mediciones realizadas a partir de las muestras de saliva. Se describen los modelos *mod.co.slv* (sin ninguna transformación en la variable respuesta ni en las variables predictoras), *mod.co.slv3* (transformando logarítmicamente la variable respuesta), y *mod.co.slv4* (transformación Box-Cox sobre la variable respuesta).

Modelo I

El modelo *mod.co.slv* se ha definido con la variable respuesta *co.post* y las variables predictoras *co.pre*, *age*, *stimulus.type*, *co.reac* y *hr.post*, tal y como se muestra a continuación:

$$Y = B_0 + B_1 (X_{co.pre}) + B_2 (X_{age}) + B_3 (X_{stimulus.type}) + B_4 (X_{co.reac}) \\ + B_5 (X_{hr.post}) + \epsilon$$

Eq. XI.E: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos de las mediciones en saliva) sin aplicar ninguna transformación en la variable respuesta ni en las covariables

Al plantear el modelo *mod.co.slv* con las variables descritas en la fórmula anterior, únicamente han resultado ser variables predictoras significativas la variable *co.pre* y *co.reac*. Por ello, se ha aplicado Akaike sobre el modelo, y éste ha determinado que las variables *stimulus.type* (no significativa, con un p-valor=0.09), y *hr.post* (no significativa con un p-valor=0.10) también se incluyan en el modelo. El modelo es significativo, y tiene un R^2 con un valor de 0.9144. Sin embargo, al aplicar los distintos test sobre los residuos del modelo, se observa que éstos no se distribuyen de manera normal y que la varianza no se distribuye de forma constante, es decir, no se cumple el supuesto de homocedasticidad. Al no cumplirse ambos supuestos, este modelo ha quedado descartado para predecir el nivel del cortisol tras aplicar un estímulo sobre el participante. Además, este modelo en comparación con los otros tres planteados, es el que ha obtenido un valor AIC y BIC para la comparación de modelos mucho más alto que los demás.

Modelo II

El modelo *mod.co.slv3* estima en un principio el valor de la variable *co.post* en función de las variables *co.pre*, *age*, *stimulus.type*, *co.reac* y *hr.post*, transformando logarítmicamente la variable respuesta.

$$\log(Y) = B_0 + B_1 (X_{co.pre}) + B_2 (X_{age}) + B_3 (X_{stimulus.type}) + B_4 (X_{co.reac}) \\ + B_5 (X_{hr.post}) + \epsilon$$

Eq. XII.E: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos de las mediciones en saliva) transformando logarítmicamente la variable respuesta co.post

Tal y como se ha explicado para el modelo I, en este caso, al conseguir únicamente un p-valor significativo en las variables *co.pre* y *co.reac*, se ha aplicado Akaike sobre el modelo, y finalmente dejando únicamente ambas variables para predecir el nivel de cortisol tras el estímulo. En este caso, el R^2 obtenido es 0.8884. A la hora de llevar a cabo el análisis de los residuos del modelo, no se ha cumplido el principio de normalidad, ya que se ha obtenido un p-valor = $2.96 \cdot 10^{-5}$, y además gráficamente se ha observado que las colas diferían del eje central. Sin embargo, el

modelo cumple el supuesto de homocedasticidad, ya que obtiene un p-valor=0.76 en el test de *ncVs* y un p-valor=0.396 en el test de *Breusch-Pagan*. En el gráfico de *Scale-Location* se observa que a simple vista también parecía que la varianza de los residuos era constante. Finalmente, sí que se observan valores *outliers*, y en referencia a la linealidad del modelo, se observa que en el gráfico *Residuals vs Fitted*, se produce una parábola, lo cual muestra la falta de linealidad del modelo. Por lo tanto, el modelo *mod.co.slv3* se descarta. Al comparar los modelos entre ellos, ha sido el modelo con un valor AIC y BIC más bajo después del modelo seleccionado (*mod.co.slv2*) y previamente explicado en el documento.

Modelo III

El último modelo que se ha planteado con los datos de las mediciones en saliva se denomina *mod.co.slv4*, y en este caso se ha aplicado la transformación Box-Cox sobre la variable respuesta *co.post*. Del mismo modo que para el biomarcador *oxitocina*, primero se ha calculado el valor de *lambda* a partir del modelo sin ninguna transformación. Se ha obtenido un valor de *lambda* = 0.50, y éste se ha aplicado sobre la variable respuesta *co.post* mediante la función $y(\lambda) = \frac{y^\lambda - 1}{\lambda}$. El modelo planteado en un en un primer momento se define mediante la siguiente fórmula:

$$\frac{Y^\lambda - 1}{\lambda} = B_0 + B_1 (X_{co.pre}) + B_2 (X_{age}) + B_3 (X_{co.reac}) + B_4 (X_{med.dos}) + B_5 (X_{gender}) + \epsilon$$

Eq. XIII.E: planteamiento inicial para el modelo que predice el nivel de cortisol post aplicación de un estímulo (utilizando la base de datos de las mediciones en saliva) aplicando la transformación Box-Cox sobre la variable respuesta co.post

Del mismo modo que para los otros modelos del conjunto de datos de la saliva, únicamente han resultado significativos las variables predictoras *co.pre* y *co.reac*, y tras aplicar Akaike, también se ha añadido la variable *stimulus.type* al modelo, ya que tiene un p-valor=0.08 (es decir, cercano al nivel de significancia de 0.05). Sin embargo, el modelo no ha aceptado la hipótesis nula de normalidad, ya que el p-valor obtenido en el test de Shapiro-Wilk tiene un valor de 0.0003, ni tampoco se ha aceptado el de la homocedasticidad, ya que se ha obtenido un p-valor menor que 0.05 en los dos tests aplicados para analizar este supuesto. Respecto a los gráficos de los residuos, se observan bastantes observaciones *outliers*, que por ejemplo afectan a la distribución de linealidad para los valores más altos, y también en la normalidad, ya que hacen que las colas de la distribución difieran del eje central. Por lo tanto, este modelo ha quedado rechazado para predecir la variable *co.post*, y además, al comparar los modelos mediante las funciones AIC y BIC, se ha obtenido un valor muy alto, descartándolo frente a los otros modelos planteados.

Anexo F

Aplicación de los modelos

En el presente anexo se muestran los resultados obtenidos tras la aplicación de la variable etnia en los modelos definidos para el biomarcador de la oxitocina y el cortisol (tanto para el modelo de la sangre como para el del serum). Se ha excluido el modelo con el conjunto de datos del cortisol completo, puesto que no ha cumplido con las suposiciones básicas para el modelo lineal.

La variable etnia es una variable categórica de cuatro niveles: *hispanic* (hispano o latino), *afroamerican* (afroamericano), *white* (blanco) u *other* (referente a las demás etnias). El software estadístico R, por defecto, transforma las variables categóricas de más de dos niveles en observaciones 0 o 1, y esto se puede observar mediante la función *model.matrix* aplicada sobre el objeto del modelo. En este caso, compara los niveles de *hispanic*, *other* y *white* con los niveles de la etnia afroamericana (que es la etnia que aparece en la primera observación del conjunto de datos *data.oxt* utilizada para definir el modelo), tal y como se observa en la Tabla V.F.

Tabla V.F: primeras seis observaciones de la transformación de la variable categórica etnia al definir el modelo de regresión en el software estadístico R obtenidas mediante la función model.matrix

	eth_hispanic	eth_other	eth_white
1	0	0	0
2	0	1	0
3	0	0	0
4	0	0	0
5	0	0	1
6	0	0	1

El resultado obtenido para el modelo de la oxitocina post aplicación de un estímulo sobre el participante se recoge en la Tabla VI.F que se muestra a continuación. En ella se puede observar que ninguno de los niveles de la etnia es significativo al 5%. Se ha aplicado Akaike sobre el modelo, para valorar la posibilidad de que aunque los diferentes niveles no fueran significativos podrían mantenerse en el modelo, pero como era de esperar con los p-valores tan elevados que se han obtenido, la variable no debe mantenerse. Además, es importante recordar que la etnia se ha asignado de forma aleatoria, por lo que es normal que no resulte significativa. Los valores del R² y del p-valor apenas difieren de los observados sin la variable etnia en el modelo.

Tabla VI.F: resultado del output obtenido tras añadir la covariante etnia en el modelo que mide el nivel de la oxitocina tras aplicar un estímulo en el participante. Se observa que la variable predictora etnia no es significativa

Predictores	Coeficiente B	Std.Err	t	Sig
<i>constante</i>	-1.30499	0.98483	-1.325	0.193055
<i>log(age)</i>	-0.60936	0.25352	-2.404	0.021220 *
<i>stimulus.type2</i>	-0.16922	0.06044	-2.800	0.007995 **
<i>log(oxt.pre)</i>	0.99318	0.07217	13.761	2.43e-16 ***
<i>log(hr.bas)</i>	0.83303	0.20942	3.978	0.000302 ***
<i>eth_hispanic</i>	0.03958	0.10979	0.361	0.720459
<i>eth_other</i>	0.04798	0.11461	0.419	0.677806
<i>eth_white</i>	0.07433	0.06914	1.075	0.289139

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

F	38.13
R ²	0.8524
p-valor	2.819e-15

En las siguientes tablas (Tabla VII.F y Tabla VIII.F) se muestra el resultado del *output* obtenido para ambos modelos del cortisol (saliva y sangre, respectivamente). En ninguno de los dos modelos se ha obtenido un p-valor significativo para los niveles de la variable etnia, y los valores del R² y del p-valor no difieren del resultado obtenido cuando la covariable etnia se elimina del modelo. Tal y como se ha observado para el biomarcador oxitocina, es normal que la variable no resulte significativa, puesto que los valores se han incluido en cada uno de los conjuntos de datos de manera aleatoria.

Tabla VII.F: resultado del output obtenido tras añadir la covariable etnia en el modelo que mide el nivel del cortisol tras aplicar un estímulo en el participante utilizando el conjunto de datos de la saliva. Se observa que la variable predictora etnia no es significativa

Predictores	Coeficiente B	Std.Err	t	Sig
constante	-0.04463	0.22350	-0.200	0.845
log(co.pre)	0.92263	0.03382	27.278	3.63e-12 ***
log(co.reac)	0.26957	0.02202	12.241	3.88e-08 ***
eth_hispanic	0.01438	0.05761	0.250	0.807
eth_other	0.15137	0.08866	1.707	0.113
eth_white	-0.04136	0.04611	-0.897	0.387
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1				

F	241.3
R ²	0.9861
p-valor	1.319e-11

Tabla VIII.F: resultado del output obtenido tras añadir la covariable etnia en el modelo que mide el nivel del cortisol tras aplicar un estímulo en el participante utilizando el conjunto de datos de la sangre. Se observa que la variable predictora etnia no es significativa

Predictores	Coeficiente B	Std.Err	t	Sig
constante	6.699e+00	9.098e-02	73.628	< 2e-16 ***
co.pre	3.927e-04	1.908e-05	20.576	< 2e-16 ***
age	4.993e-03	2.068e-03	2.414	0.023777 *
co.reac	5.137e-03	5.001e-04	10.273	2.9e-10 ***
med.dos	-2.521e-04	6.289e-05	-4.009	0.000516 ***
eth_hispanic	-3.141e-02	6.323e-02	-0.497	0.623864
eth_other	3.234e-02	7.444e-02	0.434	0.667821
eth_white	6.440e-02	3.908e-02	1.648	0.112377
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1				

F	75.02
R ²	0.9436
p-valor	9.21e-15