TFM_Metodología_Oxitocina

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

Jone Renteria

Contents

1 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.

1.1 Generación de la base de datos

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 Tas et al. (2018) y Ooishi et al. (2017). Ambos artículos 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 ha llevado a cabo un archivo Excel y posteriormente se ha cargado en R mediante la función *read_excel* del paquete *readxl*.

En total, la muestra está compuesta por 84 observaciones y 23 variables. Del total de observaciones, 32 son del artículo Tas et al. (2018) y el resto 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.

1.2 Modificacion de los datos

Antes de llevar a cabo el análisis de las variables, es necesario transformar alguna de ellas a variables categóricas y otras a variables numéricas. El resultado final se muestra a continuación:

1.3 Descriptiva de los datos

Para conocer cada una de las variables que componen el conjunto de datos, a continuación se muestra la siguiente tabla 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.

INCLUIR LA TABLA DE LAS VARIABLES

La base de datos está compuesta por dos conjuntos de datos utilizadas en 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 anterior. 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 son suficientes para generar diferentes modelos estadísticos, y aunque la variable *oxt.post* sólo se haya utilizado en uno de los dos artículos, también será suficiente, tal y como se demostrará en los siguientes apartados.

Si comparamos los datos en crudo de ambos artículos, observamos que la diferencia más significativa se da en el biomarcador cortisol. Esto es debido al método que han utilizado los dos artículos para recoger las muestras 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 Ooishi et al. (2017), en la saliva. Kaufman and Lamster (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, and 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 Peters et al. (1982) en su artículo.

En el caso de los dos artículos que se han utilizado para generar la base de datos, la diferencia más significativa se da en el caso del cortisol, por dos 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, lo primero que se ha hecho ha sido transformar el nivel de cortisol existente en el serum (sangre) en cortisol libre (igual que el medido en la saliva). Estrada-Y-Martin and

Orlander (2011) y Hammond, Smith, and 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 únicamente el 5% del cortisol en sangre es cortisol libre. Posteriormente, una vez modificadas las unidades del cortisol en sangre para que estén en las mismas unidades que en la saliva (transformar de μg/dl a pg/ml), éstos valores (tanto los valores previos al estímulo de estrés como los posteriores) se han multiplicado por 0.05, para que unicamente se tuviera en cuenta la cantidad de cortisol libre y poder compararla 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 "baja" correlación entre sangre y saliva no ha supuesto un problema. Esto es debido a los datos utilizados para generar el conjunto de datos de la oxitocina para llevar a cabo el análisis, ya que únicamente se han utilizado los valores del artículo Ooishi et al. (2017) porque como se ha comentado previamente, en el estudio de Tas et al. (2018) no han medido el valor de la oxitocina posterior al estímulo (necesaria para generar el modelo).

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.

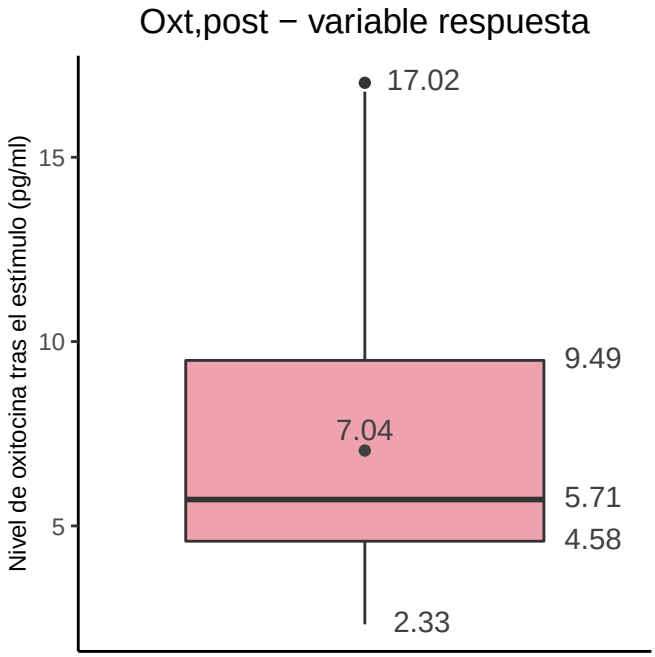
1.4 Biomarcador I: Oxitocina

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 da 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 XX anterior.

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 XXX). 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.

1.4.1 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 XX y se trata de una variable cuantitativa continua. Para obtener una descriptiva general de la variable, en la siguiente figura (Figura XX) se muestra un gráfico de cajas de esta variable:

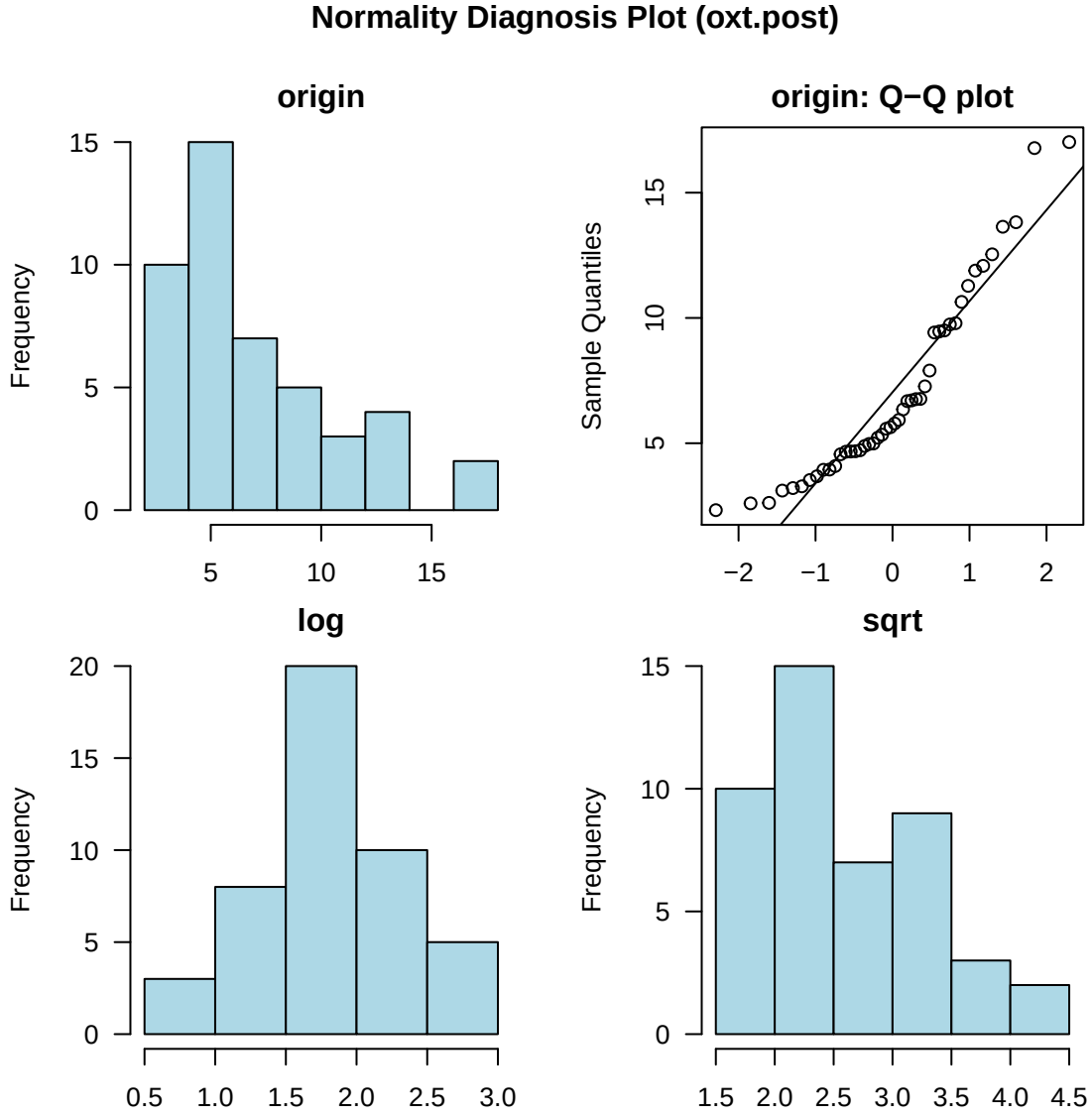


En la tabla XX se muestran los valores más significativos de la variable *oxt.post* (el valor mínimo, máximo, la mediana, la media -junto con la desviación estándar- y los cuantiles Q1 y Q3). La media de los participantes es de 7.04 pg/ml, con una desviación estándar de 3.77. En el gráfico se puede observar también un valor *outlier*, que hace referencia al valor máximo de la variable en el conjunto de datos, con un valor de 17.02 pg/ml.

METER TABLA QUE HAY EN WORD AQUI!

Aunque en el gráfico previo de cajas 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 *skewness* (el cual mide si existe simetría en la distribución de la variable) y que tiene para este caso un valor de 1.04. Los valores cercanos a cero para la observación de *skewness* se pueden considerar simétricos, y cuanto mayor sea el valor obtenido en esta columna, más difiere la variable de una distribución normal. En este caso, la variable respuesta no se aleja demasiado del valor nulo, pero en el gráfico XX se ha intuido que la variable puede estar sesgada a la derecha, debido a la distribución del tercer cuantíl del análisis. En la columna *kurtosis*, se observa el grado de presencia de valores *outliers* en la distribución, y en este caso, se obtiene un valor menor que para el caso de *skeweness*, por lo que no parece que los valores *outliers* vayan a suponer un problema en el análisis.

Es importante analizar si la variable sigue una distribución normal mediante el test de Shapiro-Wilk, fijando el nivel de significancia en un 5% y analizando el p-valor que se obtenga en el test. Este test establece como hipótesis nula la aceptación 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 *oxt.post* siga una distribución normal. Para poder observar mejor el comportamiento y la distribución de la variable, en la imagen XX que se muestra a continuación se muestra la distribución de la variable.

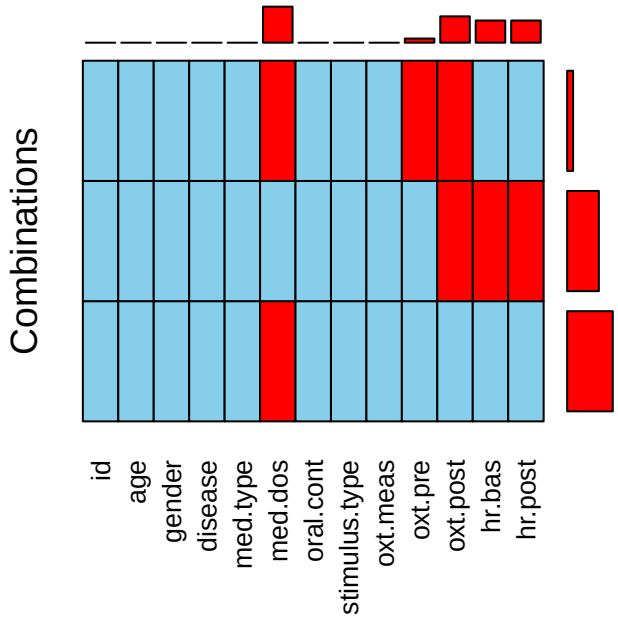


En la figura anterior se muestran cuatro gráficos: En el gráfico superior izquierdo, se muestra la distribución de la variable respuesta sin aplicar ninguna transformación sobre ella. Efectivamente, tal y como se preevía al observar el gráfico de cajas, la variable está sesgada a la derecha. En el gráfico superior derecho, también es posible observar como 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 que la variable sí estaría distribuida de forma normal, y a la derecha, la transformación de la raiz cuadrada de los datos. Por lo tanto, se concluye que parece que la variable respuesta sí que sigue una distribución normal cuando se transforma logarítmicamente. Para comprobarlo, se aplica la función *normality* sobre la variable transformada (*log(oxt.post)*), donde en este caso se obtiene un valor de p 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 *oxt.post*.

1.4.2 Valores faltantes en el conjunto de datos

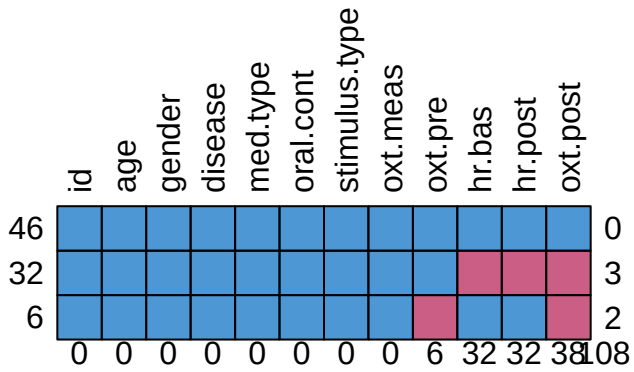
El conjunto de datos *data.oxt* está compuesto por 13 variables (incluyendo la variable respuesta (*oxt.post*) y 84 observaciones. Sin embargo, no todas las variables serán adecuadas para predecir la variable respuesta de

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, y se considera necesario analizarlo en detalle y ver en qué combinaciones y en qué situaciones se dan. Mediante la función *aggr* del paquete *VIM*, se visualiza en la figura XX la proporción de valores faltantes en el conjunto de datos (mostrados en la parte superior mediante barras), así como el gráfico de cómo son las combinaciones para los valores faltantes (en el gráfico central).



En la figura XX mostrada se observa que la variable *med.dos* es la variable que más valores faltantes incluye en el conjunto de datos. Además, para la variable respuesta *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)), y en los casos (menos frecuente) donde los valores de *oxt.pre* faltan, también lo hacen los valores de *oxt.post* (artículo Ooishi et al. (2017)). 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.

Tras eliminar la variable *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 (imagen XX), 32 tienen valores faltantes en las variables *hr.bas, hr.post* y *oxt.post*, y 6 observaciones tienen valores faltantes tanto en el nivel de oxitocina previo (*oxt.pre*) como en el posterior (*oxt.post*). Se observa de forma resumida en la figura XX que se muestra a continuación:




```
   id age gender disease med.type oral.cont stimulus.type oxt.meas oxt.pre
46  1   1      1       1        1         1             1        1       1
32  1   1      1       1        1         1             1        1       1
6   1   1      1       1        1         1             1        1       0
    0   0      0       0        0         0             0        0       6
   hr.bas hr.post oxt.post
46      1       1        1   0
32      0       0        0   3
6       1       1        0   2
       32      32       38 108
```

En la tabla XX se resumen los valores mostrados en la imagen previa.

INCLUIR TABLA HECHA A MANO EN VEZ DE LA DE R AUTOMÁTICA

Como en 32 observaciones (38.1%) hay datos faltantes para *oxt.post* y se trata de 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. Se filtran las observaciones no completas del conjunto de datos *data.oxt* mediante la función *complete.cases()*.

Antes de filtrar los datos, había 108 valores NA en total, y tras la eliminación de todos los valores faltantes, el conjunto de datos tiene hasta el momento 46 observaciones, 12 variables y ningún valor faltante.

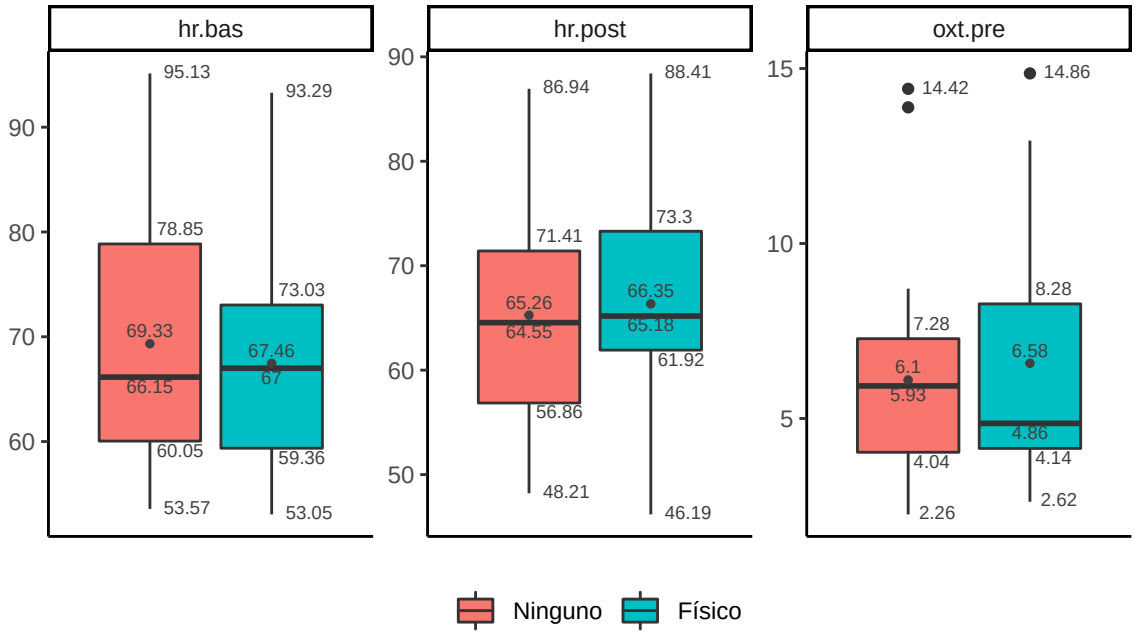
Tras analizar la base de datos vez filtradas las observaciones, se llega a la conclusión que ahora las variables categóricas *gender*, *disease* y *oxt.meas* únicamente tienen un nivel de respuesta, por lo tanto no se incluirán en los modelos que se plantearán en las siguientes secciones, puesto que no permiten la comparación con otros niveles para esa misma variable. También se elimina la variable categórica *oral.count*, puesto que ninguna participante de los estudios citados tomaba anticonceptivos orales y la variable no es por lo tanto significativa en el estudio.

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

1.4.3 Variables predictoras

De las 6 variables que componen el conjunto de datos, 5 se consideran variables predictoras, ya que la otra es la variable respuesta. Estas variables son las siguientes: *age*, *stimulus.type*, *oxt.pre*, *hr.bas* y *hr.post*, totas ellas descritas en la tabla XX. A excepción de la variable *stimulus.type*, las demás variables son cuantitativas. La variable *age*, es una variable cuantitativa discreta, pero las demás, son variables cuantitativas continuas. La variable *stimulus.type*, es una variable categórica con dos niveles: 0 para cuando no se aplica un estímulo estresante sobre la persona y 2 cuando el estímulo de estrés se aplica sobre el participante mediante música muy alta y ritmo muy acelerado. Aunque la variable *med.type* se haya eliminado del conjunto de datos final para la oxitocina, es importante destacar que todas las variables se han medido mediante muestras de saliva. Al haber únicamente una variable categórica, no es relevante reportar tablas cruzadas en este caso. Sin embargo, es sabido que hay 23 observaciones donde no se aplica un estímulo (*stimulus.type==0*) y 23cuando *stimulus.type==2*, es decir se aplica un estímulo que se basa en música de alto tempo.

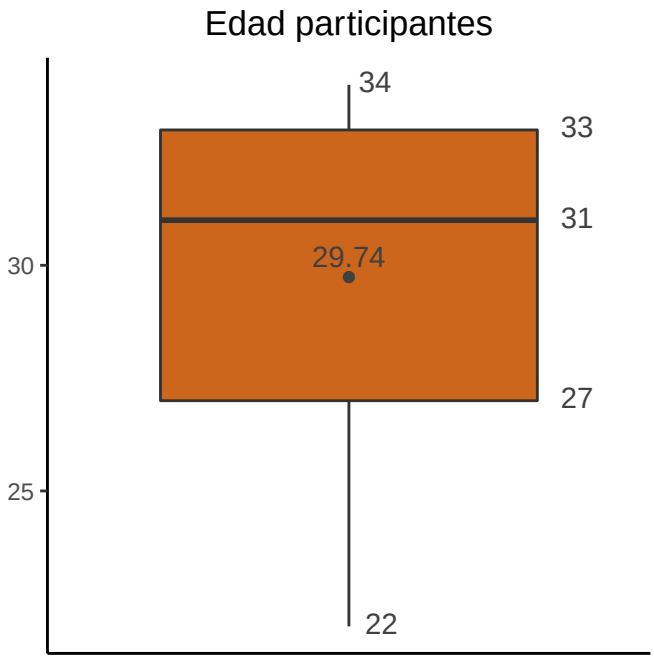
Tal y como se ha realizado para la variable respuesta, a continuación se muestra la distribución de las variables numéricas *oxt.pre*, *hr.bas* y *hr.post* según el tipo de estímulo aplicado sobre ellas:



En ninguno de los tres gráficos de la figura XX a simple vista se observa que la variable esté distribuida de forma normal. En algunos grupos (*hr.bas* sin estímulo o *oxt.pre* con y sin estímulo) parece que las variables están muy sesgadas. Para analizar los valores numéricamente, en la tabla XX que se muestra a continuación se describen los valores de las tres variables en general tanto clasificándolas por cada tipo de estímulo como en general.

METER TABLA DE WORD AQUI!

En la distribución mostrada mediante el gráfico de cajas (figura XX) 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 y entonces la distribución es igual para ambos grupos de estímulos.

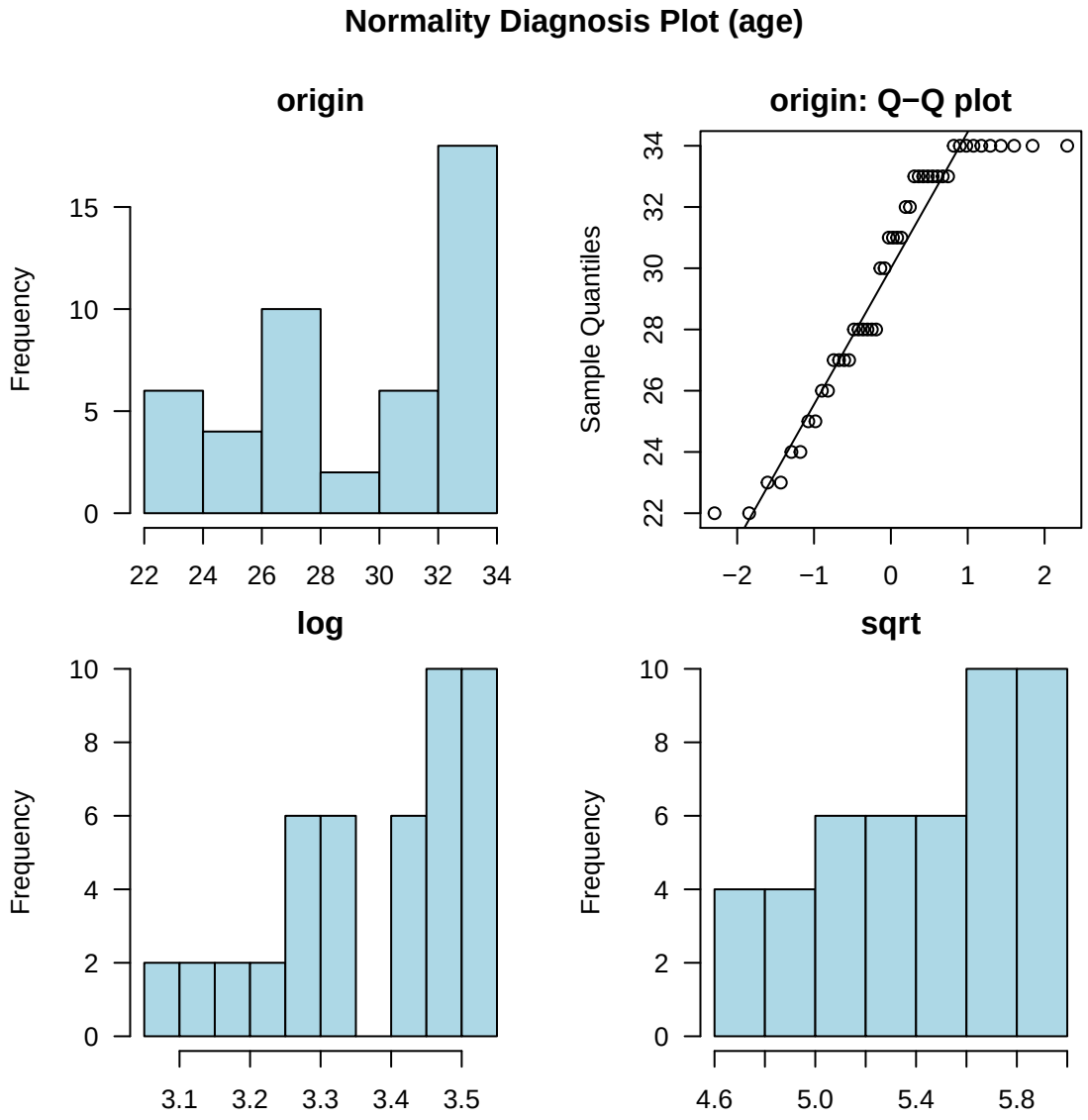


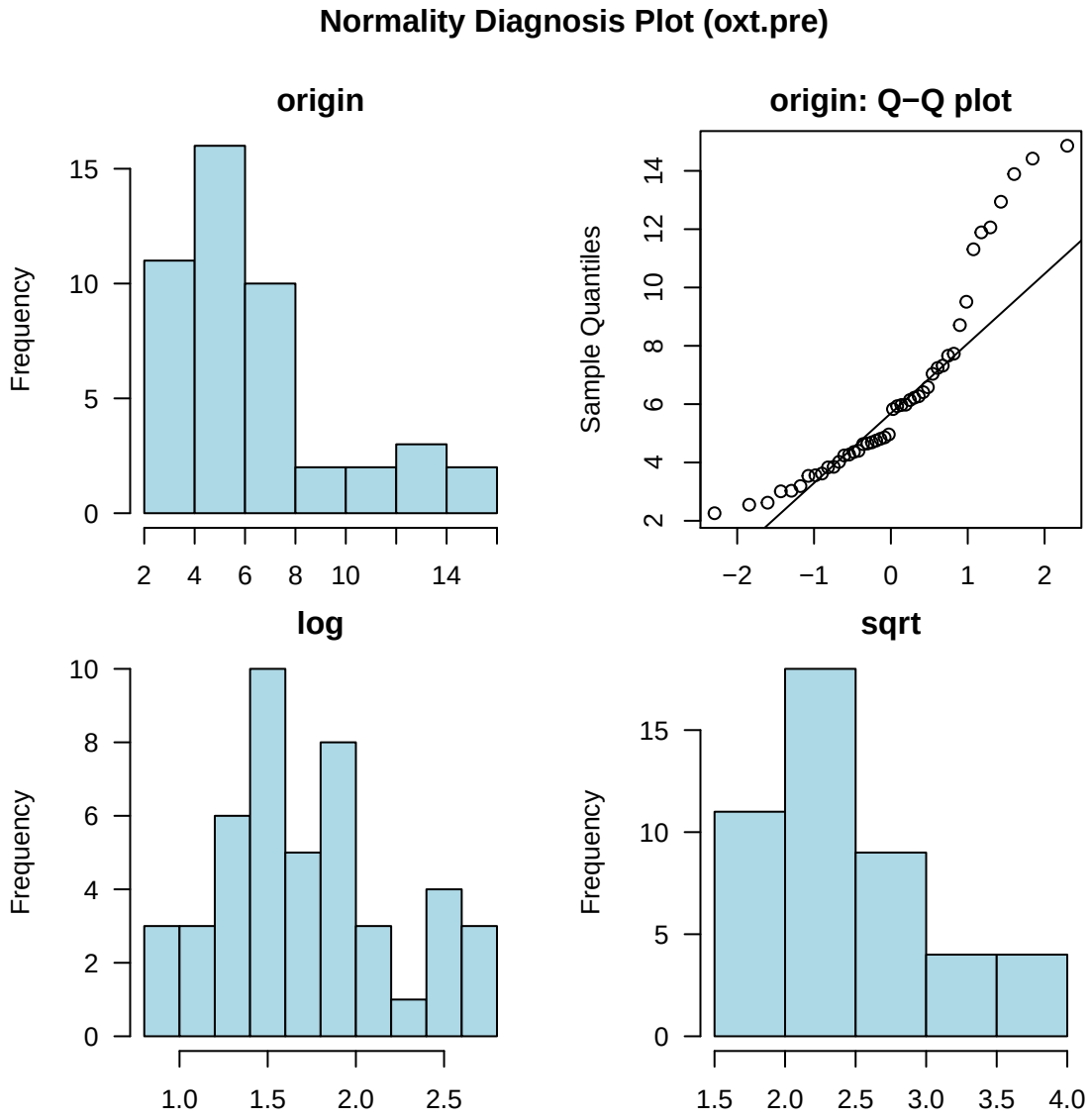
De la misma manera que con las otras variables numéricas, en la tabla XXX se muestra el resumen de los valores de la variable edad.

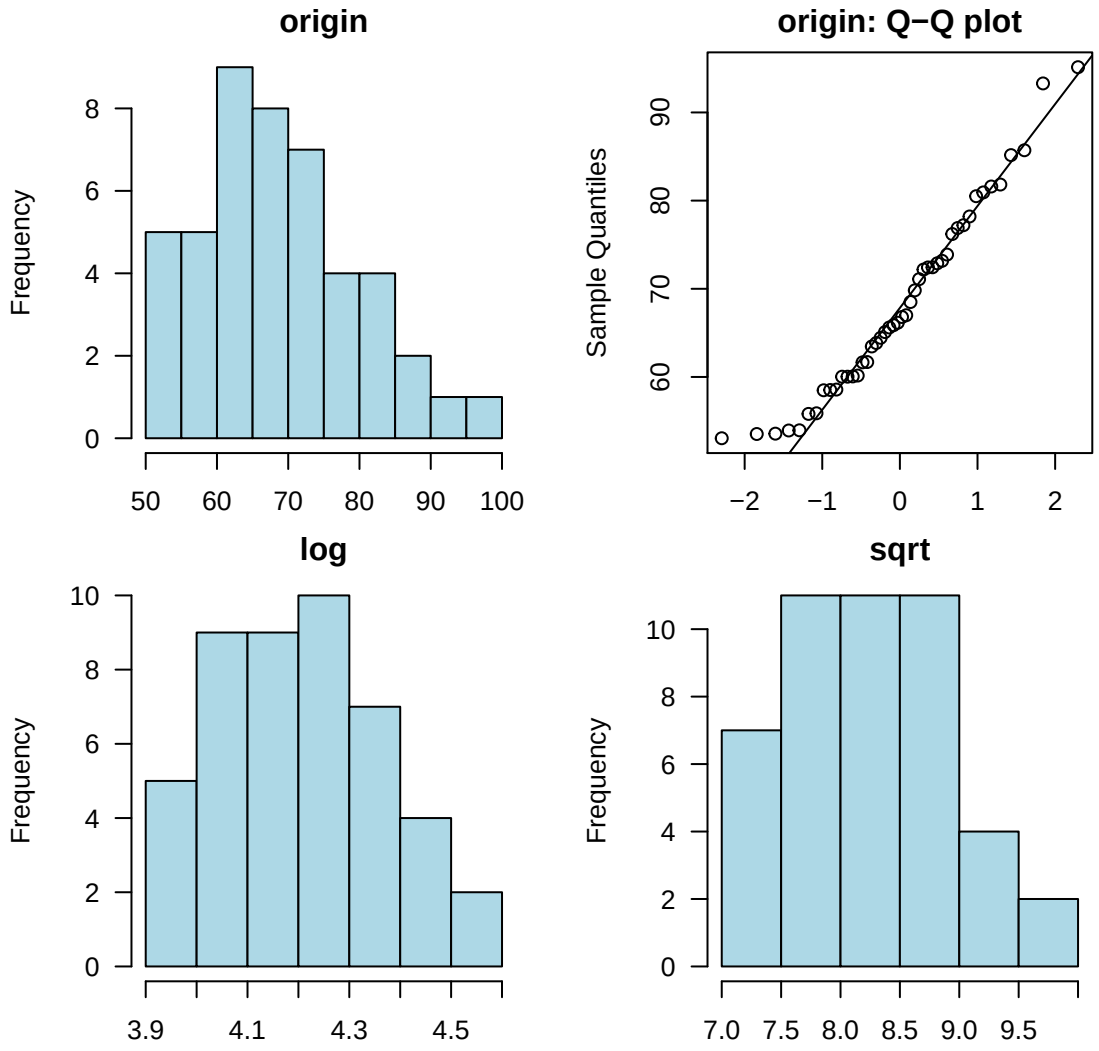
METER TABLA QUE HAY EN WORD AQUI!

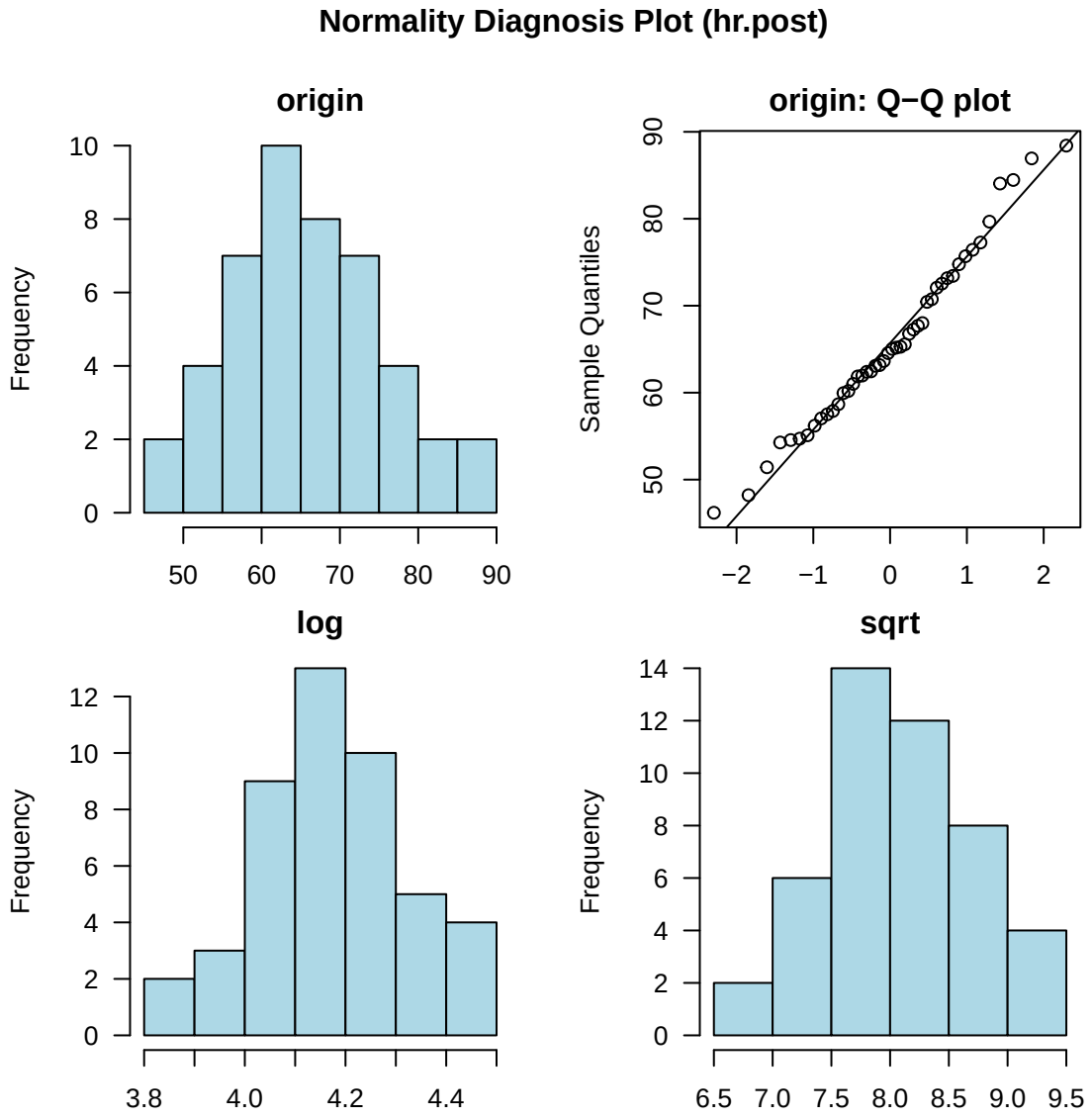
Para analizar el comportamiento general de las variables, es posible observar el valor de *skewness* para la simetría y el valor de *kurtosis* para los valores *outliers* de las variables numéricas. En este caso, la variable cuyo valor de *skewness* es más alto es *oxt.pre*, con un valor de 1.22, muy 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 también en 5%. Del test se obtiene que la variable que menos se asemeja a una distribución normal es *oxt.pre* (p-valor 5.99x10-5), seguida de la variable *age*. 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 *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 unos gráficos obtenidos a partir de la función *plot_normality* para las variables *oxt.pre*, *age*, *hr.bas* y *hr.post*.


Normality Diagnosis Plot (age)
origin
Frequency
0
5
10
15
22
24
26
28
30
32
34
origin: Q−Q plot
Sample Quantiles
22
24
26
28
30
32
34
−2
−1
0
1
2
log
Frequency
0
2
4
6
8
10
3.1
3.2
3.3
3.4
3.5
sqrt
Frequency
0
2
4
6
8
10
4.6
5.0
5.4
5.8


Normality Diagnosis Plot (oxt.pre)
origin
Frequency
origin: Q−Q plot
Sample Quantiles
log
Frequency
sqrt
Frequency


Normality Diagnosis Plot (hr.bas)
origin
Frequency
0
2
4
6
8
50
60
70
80
90
100
origin: Q−Q plot
Sample Quantiles
60
70
80
90
−2
−1
0
1
2
log
Frequency
0
2
4
6
8
10
3.9
4.1
4.3
4.5
sqrt
Frequency
0
2
4
6
8
10
7.0
7.5
8.0
8.5
9.0
9.5



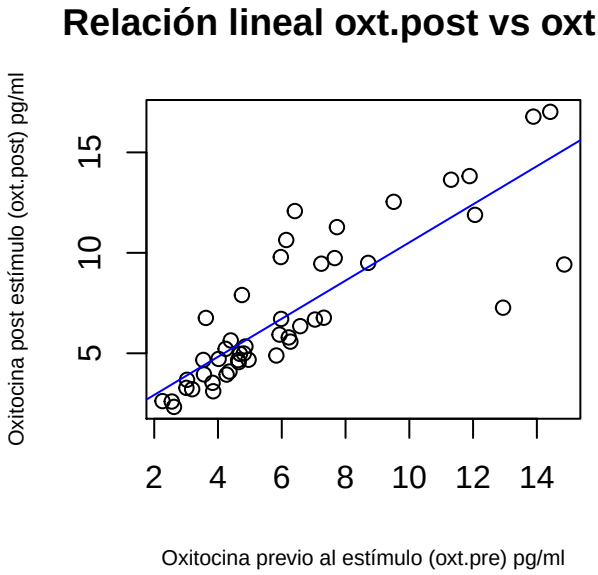
Los outputs de la función *plot_normality* para cada una de las variables numéricas mencionadas muestra que el resultado que se observa está relacionado con el p-valor analizado, ya que el histograma que muestra una distribución normal sin transformación es el de la variable *hr.post* (p valor=0.5). En la variable *hr.bas* (p-valor=0.08) se observa que la variable podría estar sesgada a la derecha, y la transformación logarítmica muestra una pequeña mejoría de la variable respecto a la original. De la variable *oxt.pre* se observa que no está distribuida de forma normal, y que la distribución puede que mejore ligeramente al transformar logarítmicamente la variable. Finalmente, la variable *age* muestra falta de normalidad a simple vista, tanto en la versión transformada como en la que no lo está. El p-valor de la variable *age* si se transformase logarítmicamente seguiría siendo muy pequeño (p-valor=0.0002) aunque superior al p-valor obtenido sin aplicar la transformación. Para las variabes *oxt.pre* y *hr.bas* transformadas logarítmicamente, sí que se obtiene un p-valor superior al 5%, y por lo tanto no hay evidencia suficiente para rechazar la hipótesis nula (p-valor 0.22 y 0.28 respectivamente). 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 sí que mejora la normalidad.

1.4.4 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 logarítmiamente, mejoran su distribución y se asemejan a una distribución normal. Por ello, como las variables podrán ser modificadas por esa razón, se aplica el método de correlación *Spearman*, en lugar del método *Pearson* que es el de defecto de la función. Aplicando este método, se evita que el coeficiente de correlación varíe en caso de que la variable sea transformada. En la siguiente tabla, se muestra la matriz de correlación de las variables del conjunto de datos:

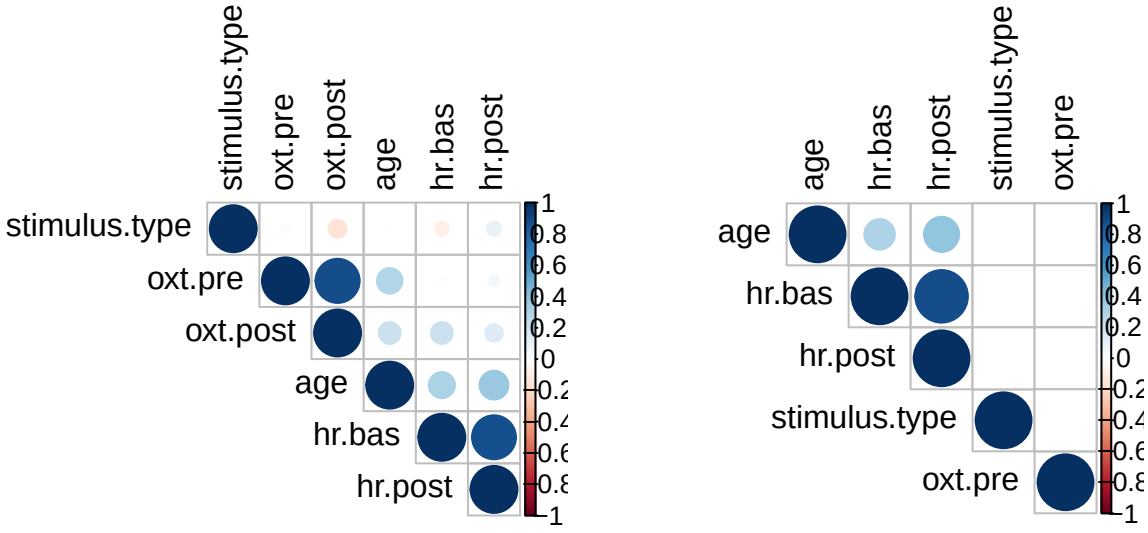
INCLUIR TABLA DE MATRIZ DE CORRELACIONES

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 predictora. En este caso, se observa que la variable *oxt.post* tiene un coeficiente de correlación de 0.885, positivo y muy alto con la variable *oxt.pre* (la correlación entre ambas se muestra en la Figura XX). Es la correlación más alta entre la variable respuesta y las variables predictoras, ya que las otras tienen coeficientes inferiores a 0.22.

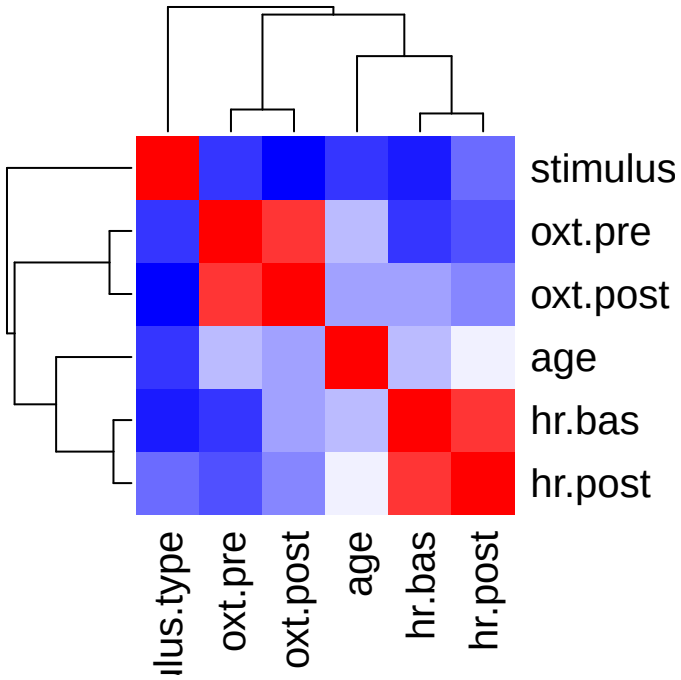


En el caso de la correlación entre las variables predictoras, en la tabla XX se observa una correlación muy alta entre ambas variables que definen el ritmo cardiaco, *hr.bas* y *hr.post*, con un coeficiente de 0.877. Esta correlación tan elevada supone que a la hora de plantear los modelos, una de ellas deba excluirse como variable predictora, para que los coeficientes que se obtengan en el modelo sean fiables. También es posible analizar la correlación entre las variables según el p-valor, y ver cuales son significativos al 5%: en este caso, se obtiene un p-valor significativo para la combinación entre ambas variables del ritmo cardiaco (p-valor: 2.22x10-16), 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 analizar la correlación entre las variables de forma gráfica, a continuación se muestra la Figura XX, obtenida a través de la función *corrplot*:



En la imagen de la izquierda, 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 como una vez más se muestra que los ritmos cardiacos están correlacionados, y en menor medida, la variable edad con ambas mediciones. También se observa correlación entre *oxt.pre* y *oxt.post*. En el gráfico de la derecha, se muestran también los coeficientes de correlación pero se han eliminado aquellos valores de las variables predictoras que no son significativos al 5%. Una vez más, la correlación se observa en la combinación de las medidas en los ritmos cardiacos y en la edad con ambas medidas. Finalmente, para concluir el análisis de la correlación, a continuación en la Figura XX se muestra un mapa de calor (*heatmap*) con los valores de la matriz de correlación mostrada previamente.



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

1.4.5 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 *oxt.post*. El modelo tiene que cumplir ciertas características, como la independencia de las variables predictoras. Sin embargo, de las 5 variables predictoras, se ha observado que dos de ellas están altamente correlacionadas, por lo que no se pueden incluir ambas en los modelos que se plantean. Para el análisis de la oxitocina, se plantea la eliminación de *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 *hr.bas* ($R^2 = 0.859$ frente a $R^2 = 0.52$ obtenido con el modelo que incluye *hr.post*). En el Anexo XX, 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 respecto al comportamiento de los residuos y además, el valor del R^2 es superior. Se trata del modelo denominado *mod.oxt2*, donde todas las variables numéricas (tanto variable respuesta como predictoras) se han transfromado logarítmicamente. Aunque en el subapartado anterior se haya observado que la variable edad no mejoraba su distribución al transformarla en logarítmica, se ha observado que el valor de R^2 ajustado es ligeramente superior transformándolo, por ello también se ha modificado junto con las otras variables numéricas.

La fórmula del modelo que se plantea es la siguiente:

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

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

COPIAR IMAGEN DEL OUTPUT DEL MODELO MOD.OXT2

```
Call:
lm(formula = log(oxt.post) ~ log(age) + stimulus.type + log(oxt.pre) +
```

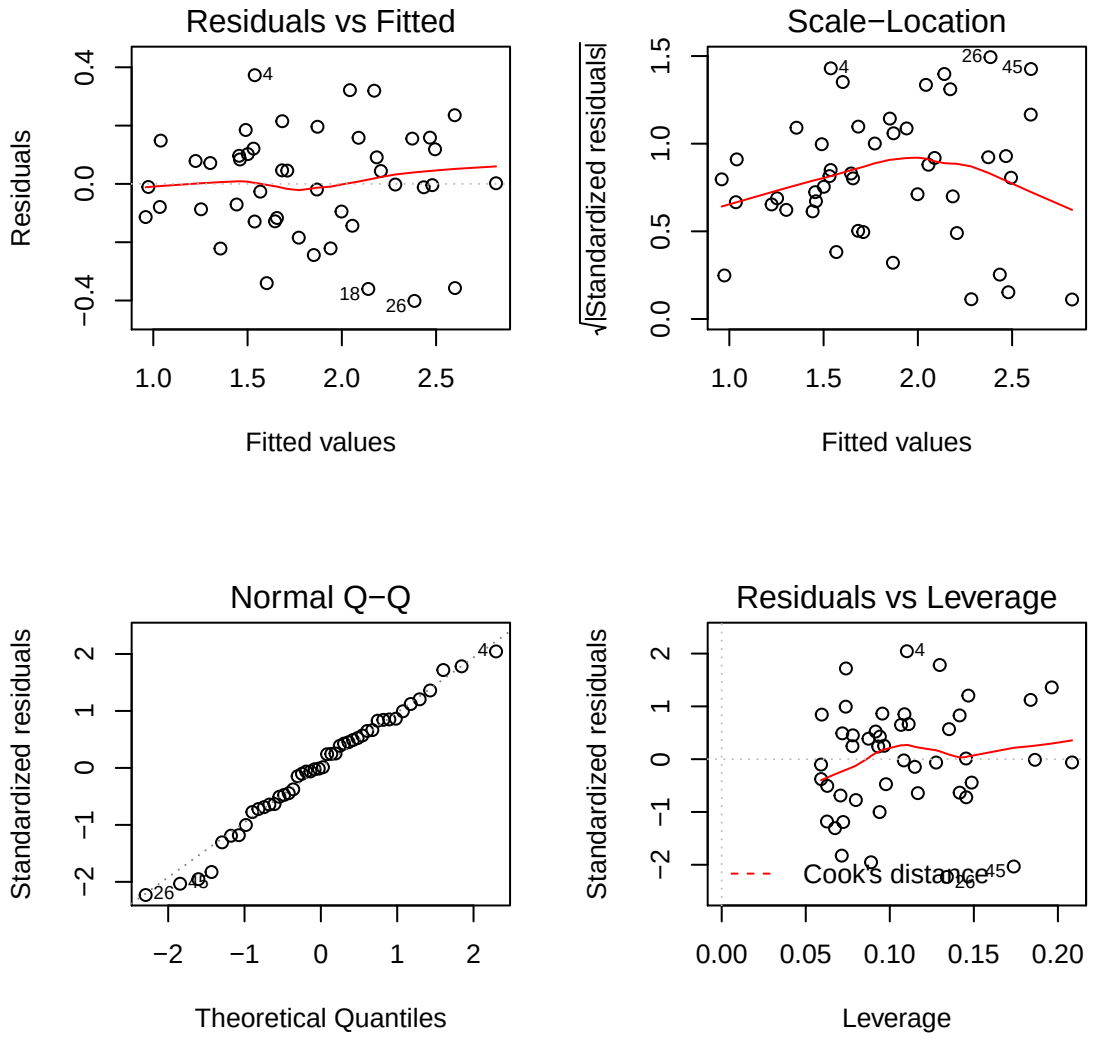
```
    log(hr.bas), data = data.oxt)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4014 -0.1161  0.0000  0.1204  0.3729

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -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 ***
log(hr.bas)     0.84390    0.20285   4.160 0.000158 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1933 on 41 degrees of freedom
Multiple R-squared:  0.8716,    Adjusted R-squared:  0.859
F-statistic: 69.56 on 4 and 41 DF,  p-value: < 2.2e-16
```

En la figura 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 los errores para las siguientes observaciones. A continuación, en la figura XX, se muestran cuatro gráficos diferentes que describen los residuos del modelo *mod.oxt2*.



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

- Linearidad: 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 casi sobrepuesta 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 si que se encuentran sobrepuestos.
- 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 termino 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 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. En este caso, se observa que hay algún valor *outlier* pero ninguno de ellos está separado por la distancia de Cook.

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*.

- **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 si 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 del mismo nombre. 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.

- **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 de 0.14 y 0.59 respectivamente, y por ello se acepta que la varianza de los residus es constante, y con ello 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 (necesario para el test de *Bartlett*), 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 por cada grupo de la variable *stimulus.type*, cuando debería haber mínimo dos grupos por cada observación para poder aplicar el test correctamente.

- **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 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 destacar que en el diseño del modelo se ha eliminado la variable *hr.post* puesto que estaba altamente correlacionada con *hr.bas*.

- **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 (mímino 1.01 y máximo 1.25), por lo tanto *cercanas* a cero y con ello, suficiente para rechazar el principio de multicolinealidad en los residuos del modelo *mod.oxt2*.

1.4.6 Conclusión modelos

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 el documento) y *mod.oxt*, *mod.oxt3*, y *mod.oxt4* descritos en el Anexo X, 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 mejores resultados se ha obtenido. 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 X) 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 em 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* (AIC=-13.94 y BIC=-2.97) 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 el modelo que mejor predice los datos de la oxitocina tras un estímulo es el modelo *mod.oxt2*, ya que, tal y como se ha mencionado previamente, aparte de ser el modelo que mejor cumple con las suposiciones de los residuos para un modelo lineal, es el que tiene un valor ajustado de R^2 más alto. Además, en comparación con los otros modelos, es el modelo que mejor ajuste muestra para los valores observados. Por lo tanto, 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 es el modelo *mod.oxt2*. La ecuación del modelo es la siguiente:

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

Siendo cada término,

- $log(Y)$: variable respuesta *oxt.post* transformada a logarítmica.
- -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 Anexos - BIOMARCADOR OXITOCINA (EN EL TFM: ANEXO B)

2.1 Anexo B

En el presente Anexo A 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.

2.1.1 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 suficientemente buenos como para considerarlo adecuado al predecir el nivel de oxitocina. El modelo se plantea de la siguiente manera:

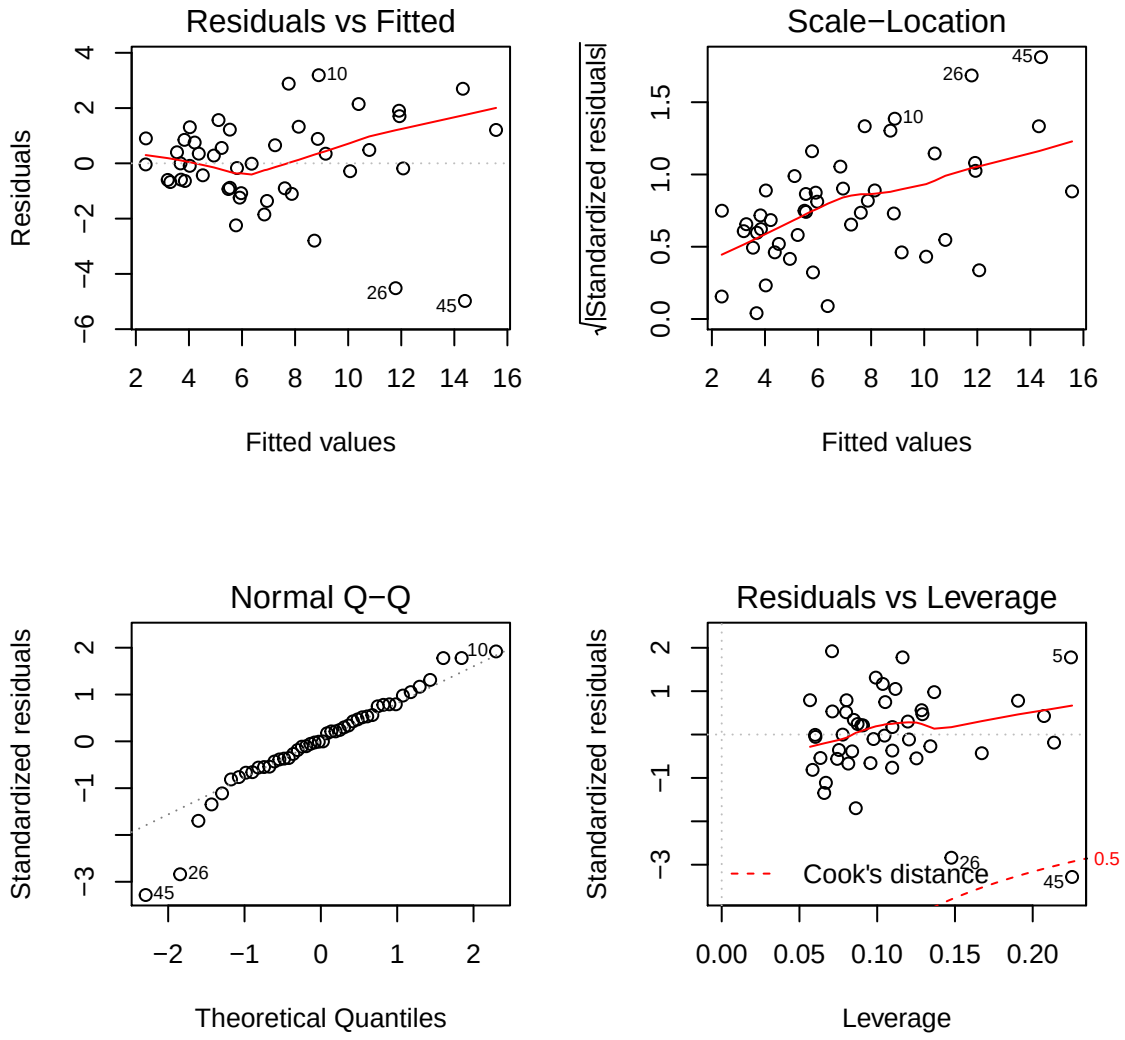
$$Y = B_0 + B_1\ (X_{age}) + B_2\ (X_{stimulus.type}) + B_3\ (X_{oxt.pre}) + B_4\ (X_{hr.bas}) + \epsilon$$

Tras su definición en R, el resultado obtenido del sumario del modelo se muestra en la Figura XX mostrada a continuación:

INCLUIR IMAGEN, mod.oxt_summary_output_ANEXO

Del resumen obtenido mediante la función *summary* del modelo planteado observamos 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 R^2 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:



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

- Linearidad: 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 sobrepuesta en 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 termino 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 un valor (el 45) separado 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 del mismo nombre. 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 XX 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 linearidad observada en los gráficos de los residuos mejore.

2.1.2 Modelo II

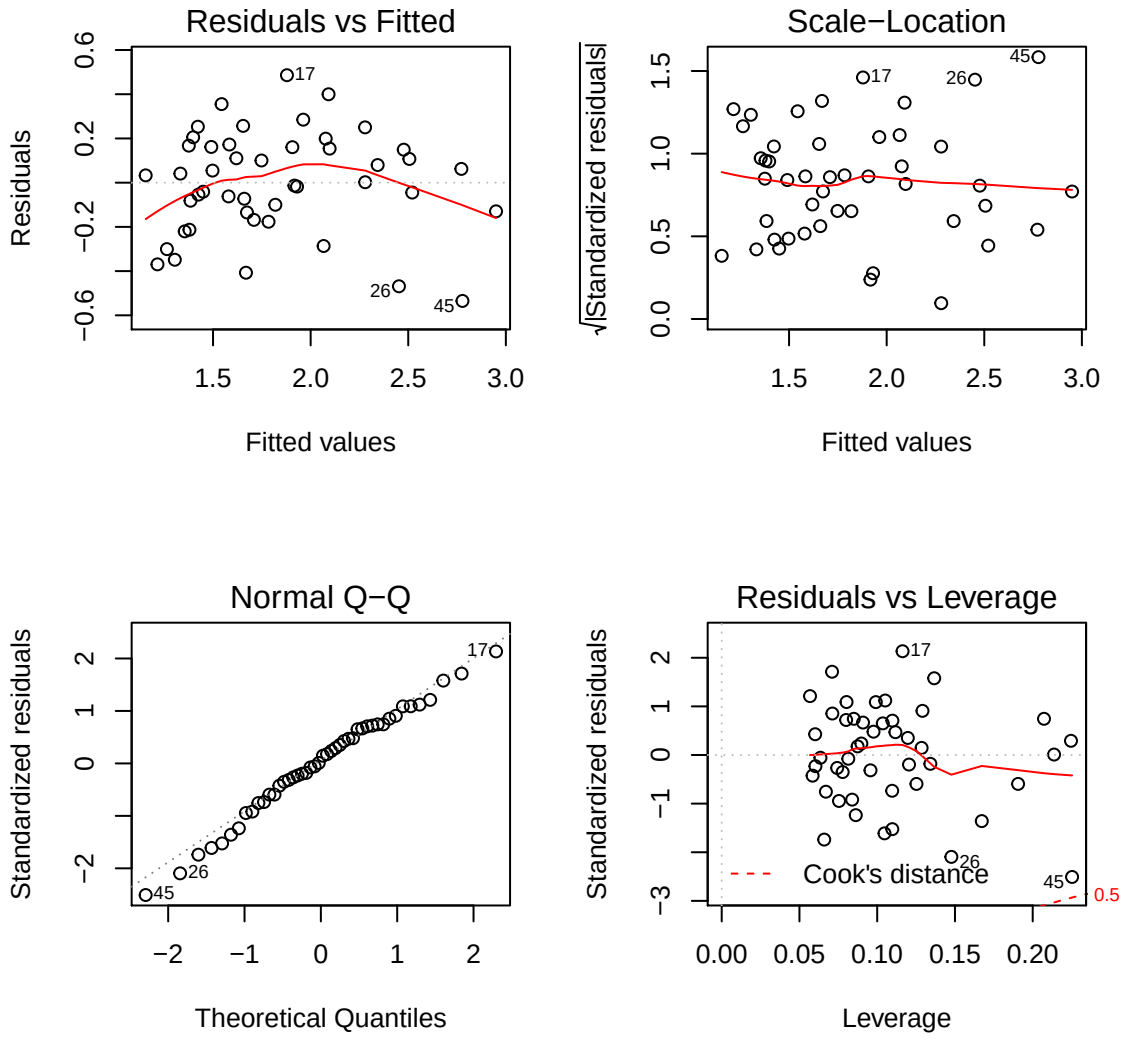
El siguiente modelo que se plantea es el modelo *mod.oxt3*, donde únicamente se modifica la variable respuesta (*oxt.post*), transformándola en una variable logarítmica. El modelo se denomina *mod.oxt3* y su planteamiento se muestra a continuación:

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

Tras aplicarlo en R, el resultado obtenido del sumario del modelo, se muestra en la Figura XX.

INCLUIR IMAGEN, mod.oxt3_summary_output_ANEXO

Tal y como se muestra en la figura XX con el *summary* 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 si 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 XX 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 XX.



A simple vista, se observa que la linearidad no se cumple ya que la línea roja no es horizontal y no está sobrepuesta en la línea central. Respecto a la normalidad, una vez más las colas parece 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 XX 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 residos 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 linearidad mostrado (*Residuals vs Fitted*) en la Figura XX 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.

2.1.3 Modelo III

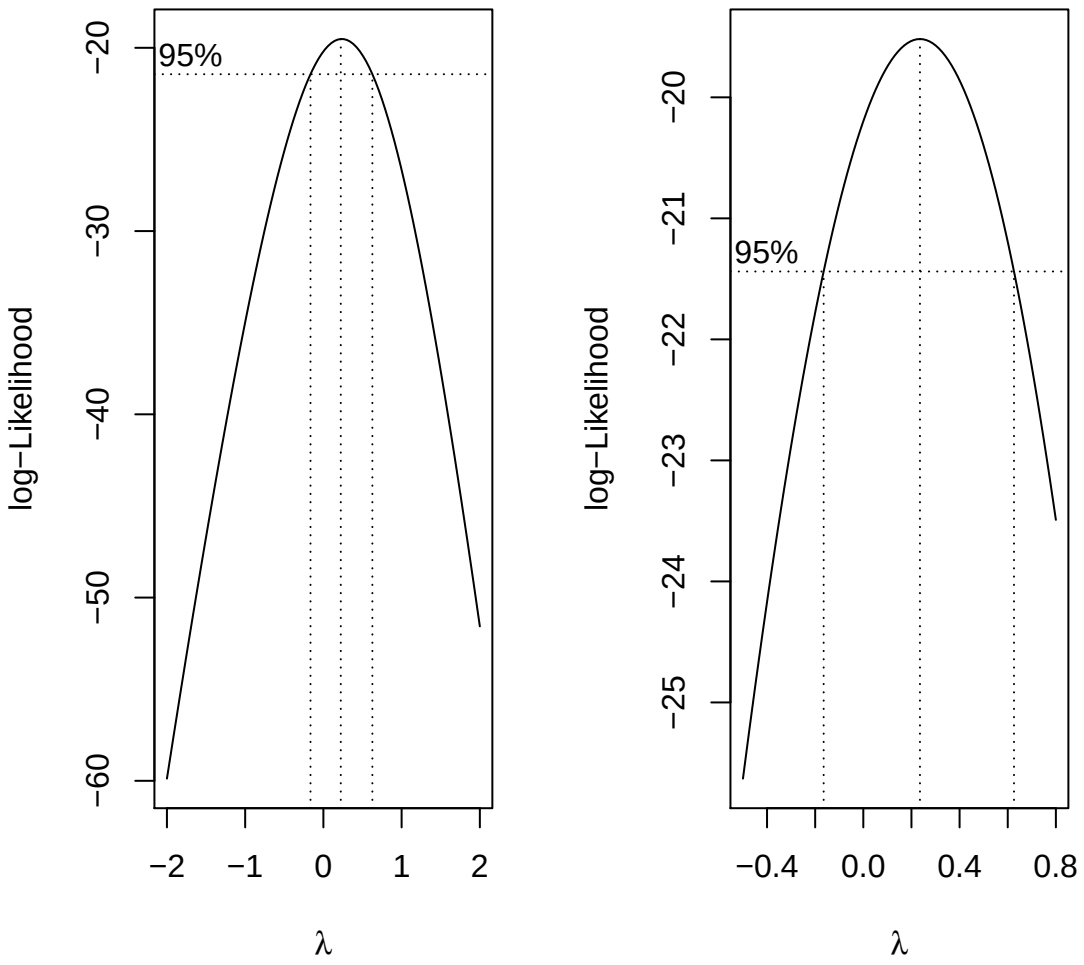
La siguiente transformación que se muestra es la transformación Box-Cox aplicada sobre la variable respuesta *oxt.post*. La transformación de Box-Cox se suele aplicar para que los residuos del modelo se asemejen a una distribución normal y también para mejorar la linearidad de los residuos. Se ha observado 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 linearidad.

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:

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

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 XX.



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:

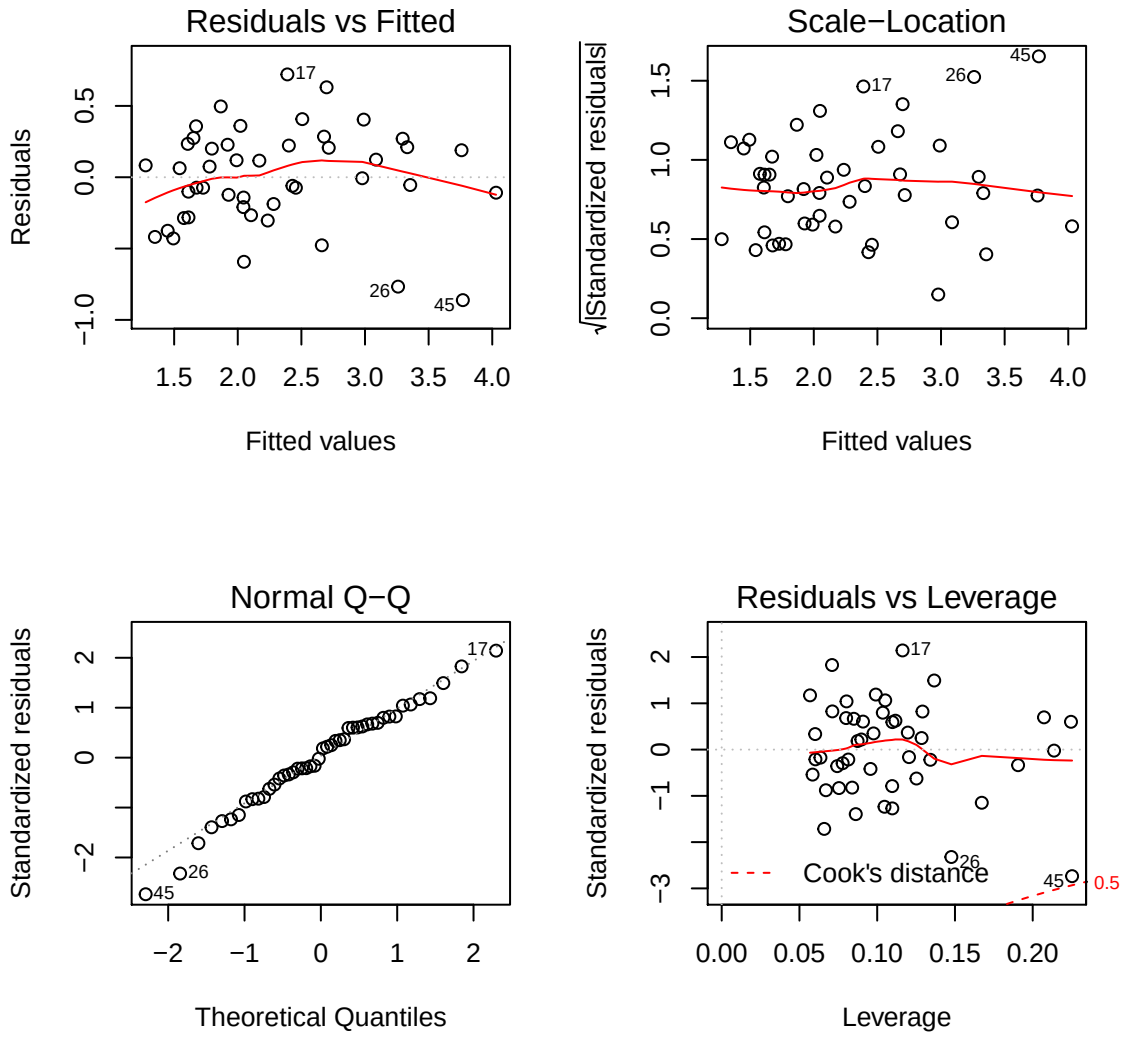
$$\frac{Y^{\lambda} - 1}{\lambda} = B_0 + B_1\ (X_{age}) + B_2\ (X_{stimulus.type}) + B_3\ (X_{oxt.pre}) + B_4\ (X_{hr.bas}) + \epsilon$$

El *output* del resumen obtenido en R tras aplicar la formula se muestra en la figura XX:

INCLUIR FOTO CON EL OUTPUT:: mod.oxt4_summary_output_ANEXO

En la Figura XX 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 XX.



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 linearidad, 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 colinearidad. 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.

References

Estrada-Y-Martin, Rosa M, and Philip R Orlander. 2011. "Salivary Cortisol Can Replace Free Serum Cortisol Measurements in Patients with Septic Shock." *Chest* 140 (5): 1216–22.

Hammond, GL, CL Smith, and DA Underhill. 1991. "Molecular Studies of Corticosteroid Binding Globulin Structure, Biosynthesis and Function." *The Journal of Steroid Biochemistry and Molecular Biology* 40 (4-6): 755–62.

Kaufman, Eliaz, and Ira B Lamster. 2002. "The Diagnostic Applications of Saliva—a Review." *Critical Reviews in Oral Biology & Medicine* 13 (2): 197–212.

McCullough, Michael E, Patricia Smith Churchland, and 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.

Ooishi, Yuuki, Hideo Mukai, Ken Watanabe, Suguru Kawato, and 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.

Peters, JR, RF Walker, D RIAD-FAHMY And, and R Hall. 1982. "Salivary Cortisol Assays for Assessing Pituitary-Adrenal Reserve." *Clinical Endocrinology* 17 (6): 583–92.

Tas, Cumhur, Elliot C Brown, Gokcer Eskikurt, Sezen Irmak, Orkun Aydın, Aysen Esen-Danaci, and Martin Brüne. 2018. "Cortisol Response to Stress in Schizophrenia: Associations with Oxytocin, Social Support and Social Functioning." *Psychiatry Research* 270: 1047–52.