# TFM_Metodología_Cortisol

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

Jone Renteria

## Contents



# 1 Cortisol analysis

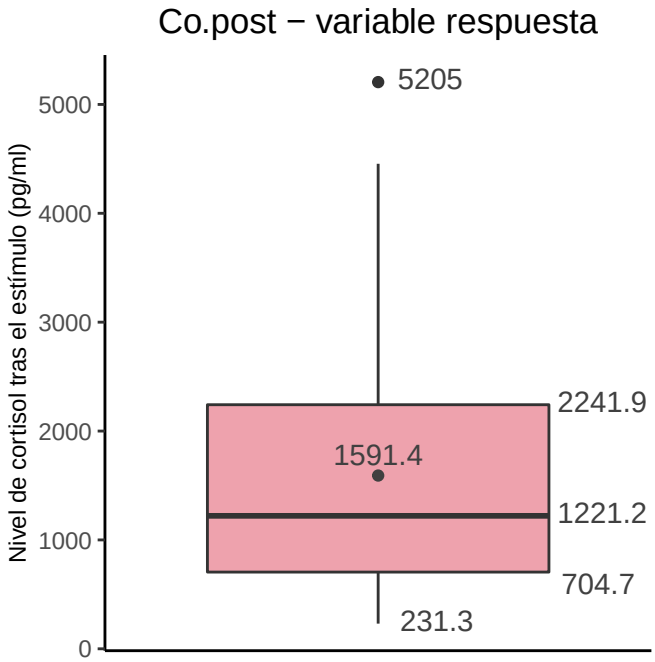
## 1.1 Biomarcador II: Cortisol

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 da 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 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 a la hora de plantear el modelo.

### 1.1.1 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 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 que describe su comportamiento:



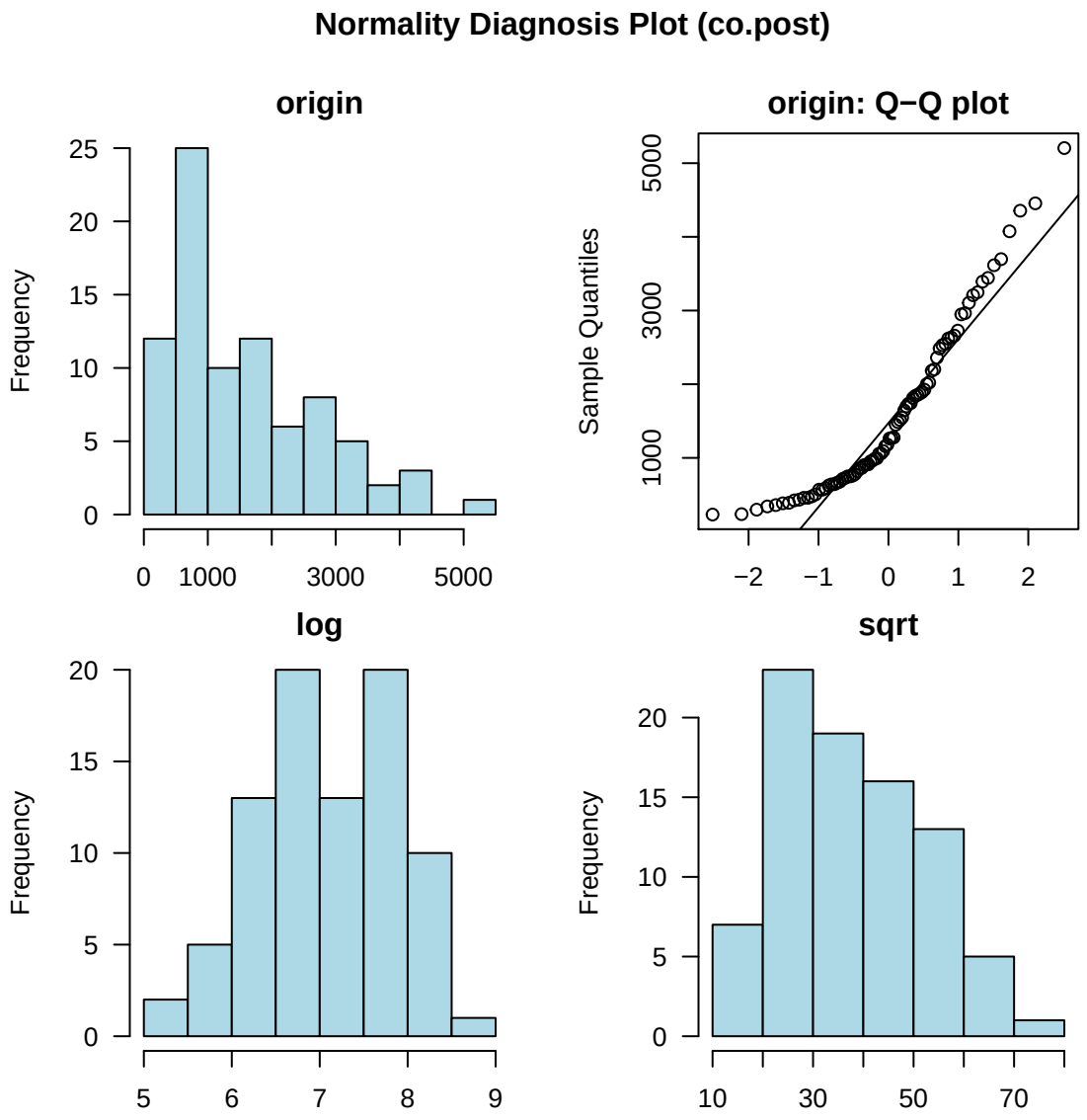

En el gráfico se observa que la variable respuesta podría estar sesgada, y que tiene un valor *outlier* (influyente),que hace referencia al valor máximo de la variable en el conjunto de datos, con un valor de 5205.0 pg/ml, tal y como se observa en la siguiente tabla (Tabla XXX). Además de este valor, en la tabla se recogen otros valores significativos de la variable *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.

**METER TABLA1 para el cortisol QUE HAY EN WORD AQUI!**

Tal y como se ha hecho 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 es posible determinar que la variable esté distribuida de manera normal, aunque esto se deberá analizar mediante diferentes tests que se llevarán a cabo posteriormente. Del gráfico en la figura XXX, se puede intuir que la variable esté ligeramente sesgada a la derecha, debido a la distribución del tercer cuantil. El valor outlier observado en la figura anterior no parece que vaya a suponer un problema, puesto que el 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

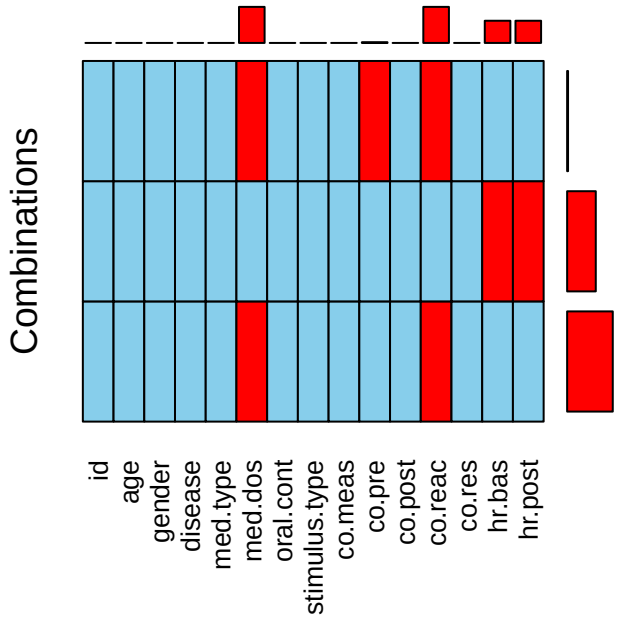
test acepta la distribución normal de los datos. En este caso, para la variable respuesta *co.post*, se obtiene un p-valor singificativo ($6.19 \text{x} 10^{-6}$), por lo tanto existe evidencia suficiente para no aceptar la hipótesis nula y considerar que la variable no sigue una distribución normal. El comportamiento de la variable se observa de forma gráfica en la siguiente imagen (figura xxx), 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 sobrepuestas en la línea contínua 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 si que se aceptaría la distribución normal de los datos.




### 1.1.2 Valores faltantes en el conjunto de datos

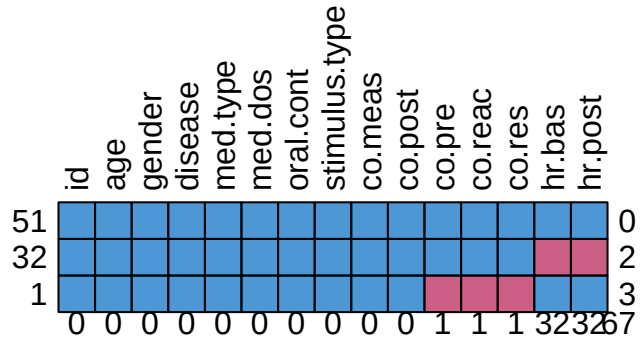
El conjunto de datos *data.co* está compuesto por 15 variables (incluyendo la variable respuesta (*co.post*) y 84 observaciones. Sin embargo, algunas variables presentan muchos valores faltantes (NA) en sus observaciones

y eto podrá suponer un problema a la hora de plantear los modelos. Mediante la función *aggr* del paquete *VIM*, se visualiza en la figura XXX 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).




En la figura xxx, se muestra que una gran proporción de valores faltantes se encuentran en las variables *med.dos*, *co.reac* y *co.res*. Sin embargo, para las dos primeras variables es posible imputar los *missing*: en el caso de la variable *med.dos*, para las observaciones donde los pacientes no tienen medicación ($med.type = 0$), se puede imputar que la dósis será por lo tanto cero. En el caso de la variable *co.reac* (índice de reaccion al cortisol, %), ésta 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 la variable *co.pre* y *co.post* (*co.post* - *co.pre*), y posteriormente calculan el porcentage 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*, también 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 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 a *co.pre* 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 xxx:




```
   id age gender disease med.type med.dos oral.cont stimulus.type co.meas
51  1   1      1       1        1       1         1              1       1
32  1   1      1       1        1       1         1              1       1
1   1   1      1       1        1       1         1              1       1
    0   0      0       0        0       0         0              0       0
   co.post co.pre co.reac co.res hr.bas hr.post
51       1      1       1      1      1       1  0
32       1      1       1      1      0       0  2
1        1      0       0      0      1       1  3
         0      1       1      1     32      32 67
```

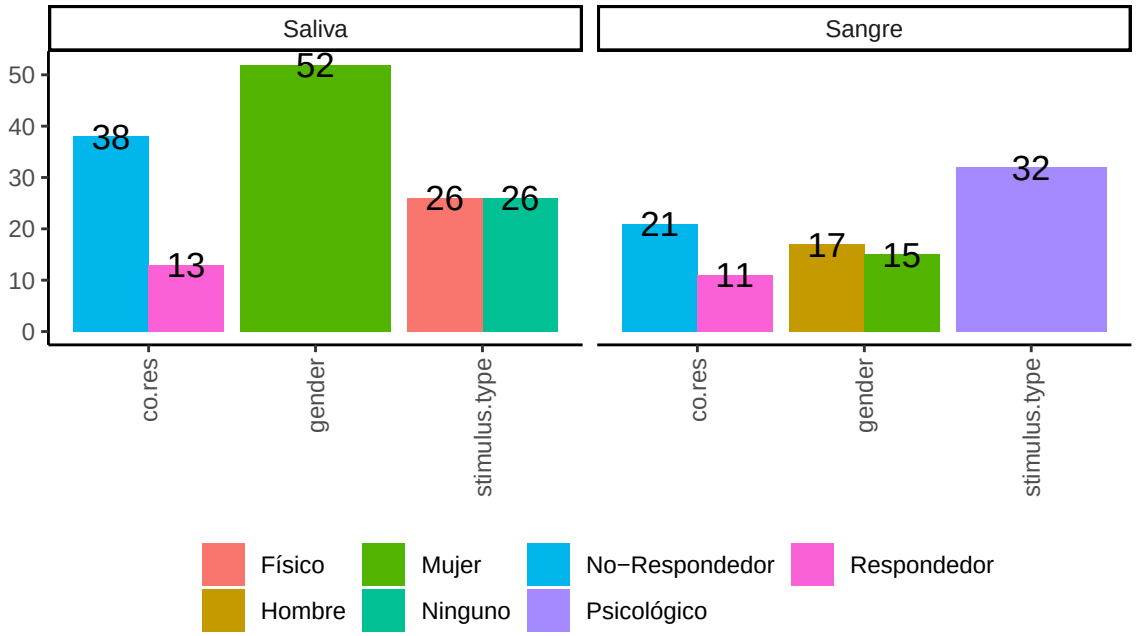
Se observa que de las 84 observaciones del conjunto de datos, 51 son observaciones completas, en 32 observaciones únicamente faltan las observaciones de las variables *hr.bas* y *hr.post*, y en una única observación falta la variable del cortisol previo al estímulo, y por lo tanto también faltan los valores en las variables *co.reac* y *co.res*. En las siguientes subsecciones (cuando se planteen los modelos y para el diseño de cada uno de ellos), se observará si se deberán eliminar las 32 observaciones donde faltan las variables *hr.bas* y *hr.post*, pero de momento se mantienen en el conjunto de datos *data.co* con un total de 84 observaciones y 15 variables.

En este caso, a diferencia del análisis de la oxitocina, las variables categóricas *gender*, *disease*, *med.type*, *stimulus.type* y *co.meas* tienen más de un nivel, por lo que todavía se mantienen en el conjunto de datos. Sin embargo, la variable *oral.count* 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 ningùn 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.

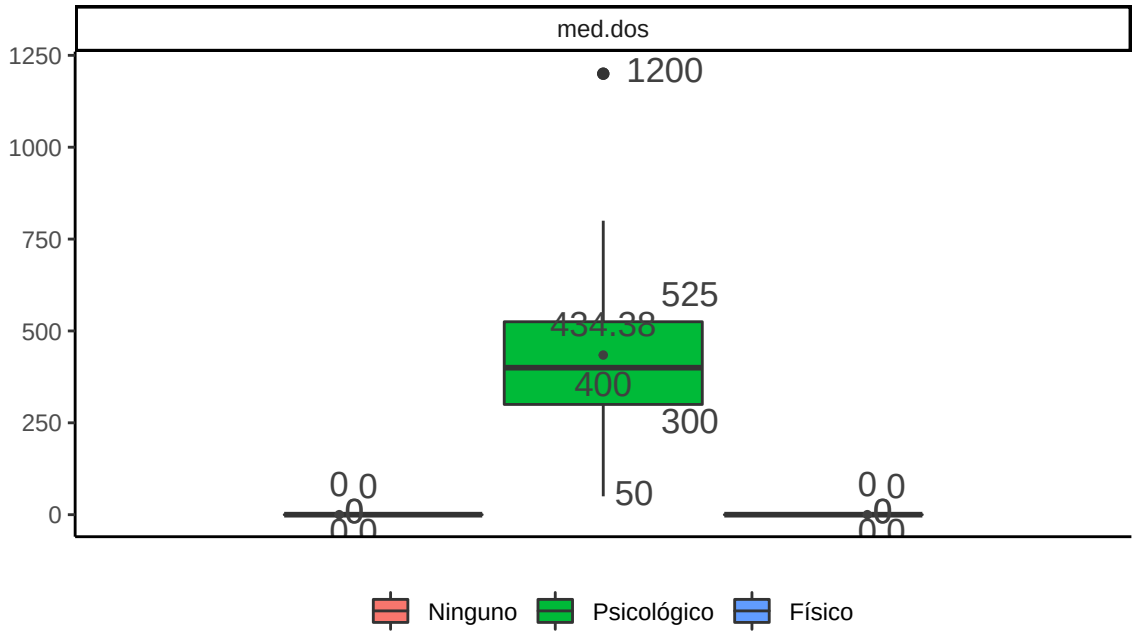
### 1.1.3 Variables predictoras

De las 13 variables que componen el conjunto de datos, 12 serán variables predictoras en los modelos que se plantearán, ya que la última es la variable respuesta. Algunas de las variables son variables numéricas (*age*, *med.dos*, *co.pre*, *co.reac*, *hr.bas* y *hr.post*) y el resto son variables categóricas (explicadas en la tabla 3 XXX).

Entre las variables categóricas, todas son factores de dos niveles, a excepción de la variable *stimulus.type*, que tiene tres (cada uno de ellos explicado en la tabla 4 XXX del documento).




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

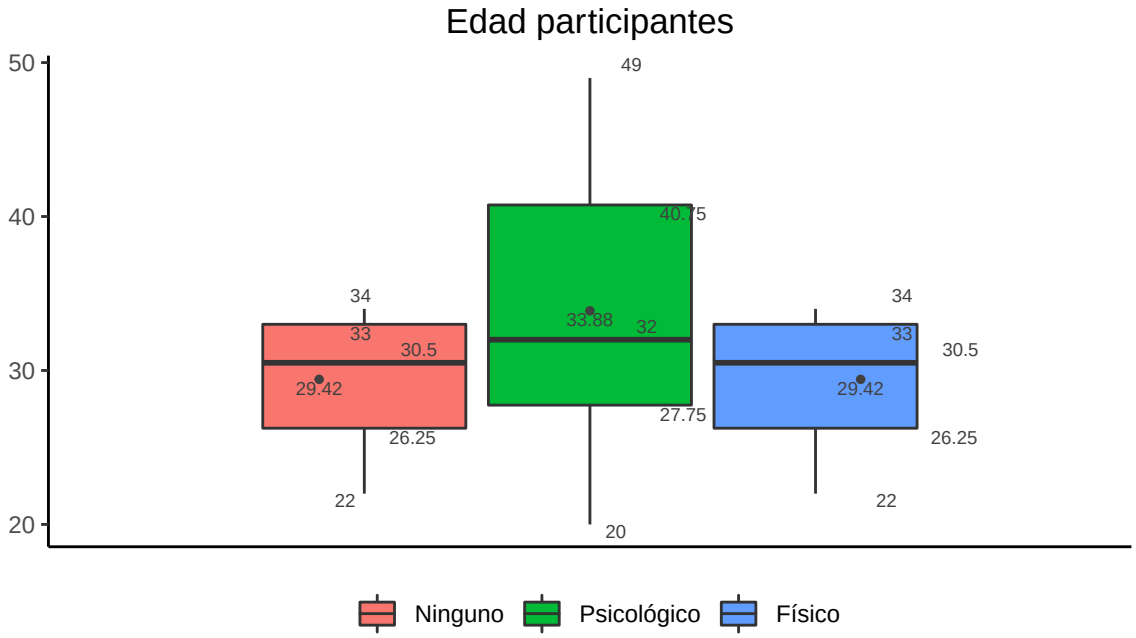



Como se ha mostrado en la tabla xxx del subapartado anterior respecto a los valores faltantes, no todas las variables tienen datos en cada tipo de estímulo. Es decir, como se muestra en la figura XXX, las variables *hr.bas* y *hr.post* por ejemplo no se calculan para el tipo de estímulo de la entrevista de trabajo, y por ello solo aparecen dos gráficos de cadas en ellos. De la misma manera, la dósis de medicación (*med.dos*) solo se mide para el tipo de estímmulo de la simulación de la entrevista de trabajo 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*, si que se han medido para los tres tipos de estímulos (en el caso de *co.reac* imputando los valores al aplicar la fórmula del modelo original) y por ello aparecen los tres gráficos de cajas para ellos. En la siguiente tabla XX 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 XXX se muestra a continuación:

**METER TABLA QUE HAY EN WORD AQUI!**

Tal y como se ha explicado en el EDA de la variable oxitocina, la variable numérica *age* es discreta y la muestra utilizada para llevar a cabo este estudio, utiliza los mismos sujetos para el tipo de estímulo *ninguno* (*stmulus.type=0*) y *música alto tempo* (*stimulus.type =2*). Se muestra a continuación en la figura XXX, donde se puede observar que las cajas para ambos estímulos son iguales.



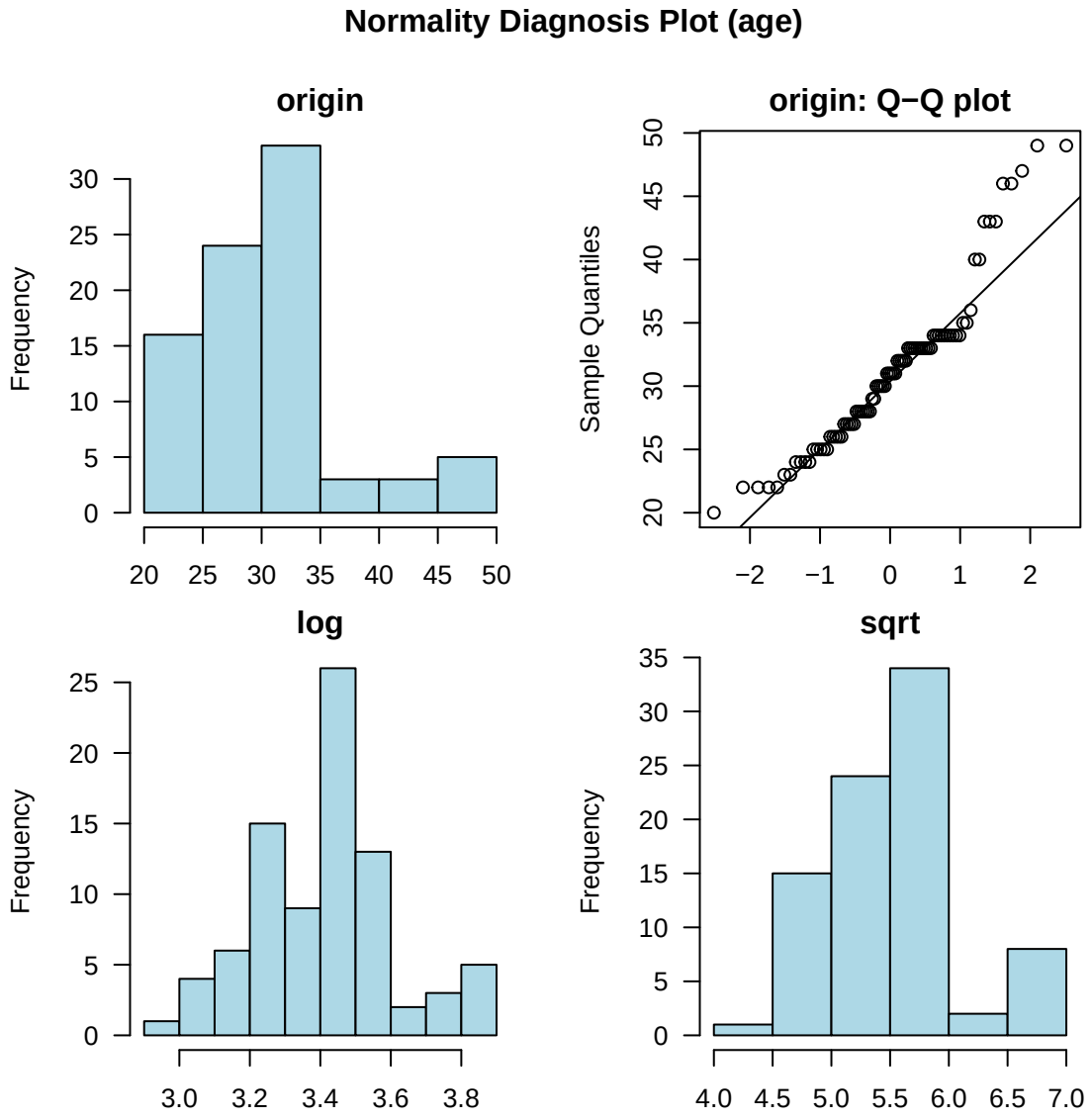

De la misma manera que con las otras variables numéricas, en la tabla XXX se muestra el resumen de los valores de la variable edad, tanto de forma general como separada por los grupos mostrados en el gráfico de cajas anterior.

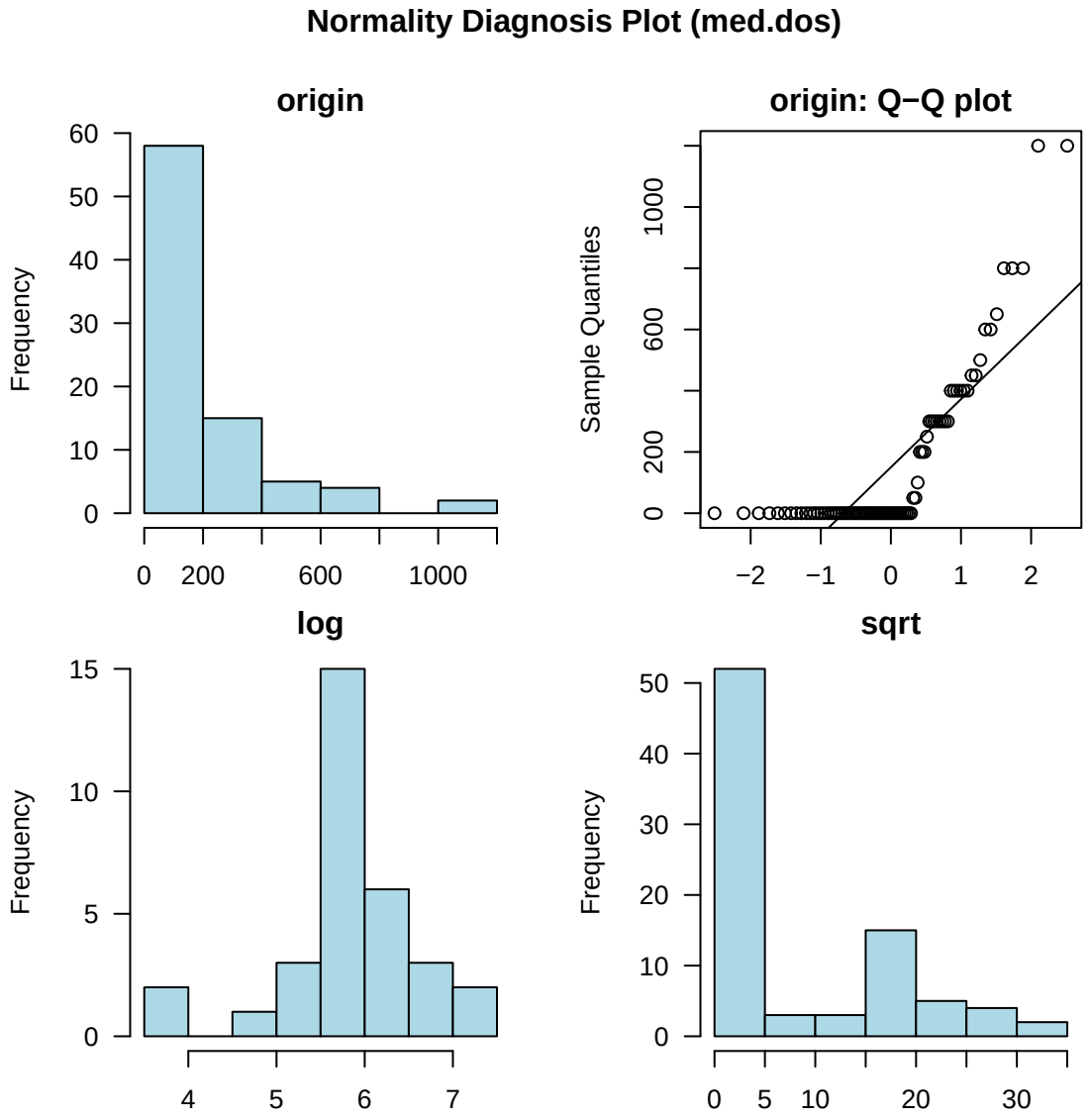
**METER TABLA QUE HAY EN WORD AQUI!**

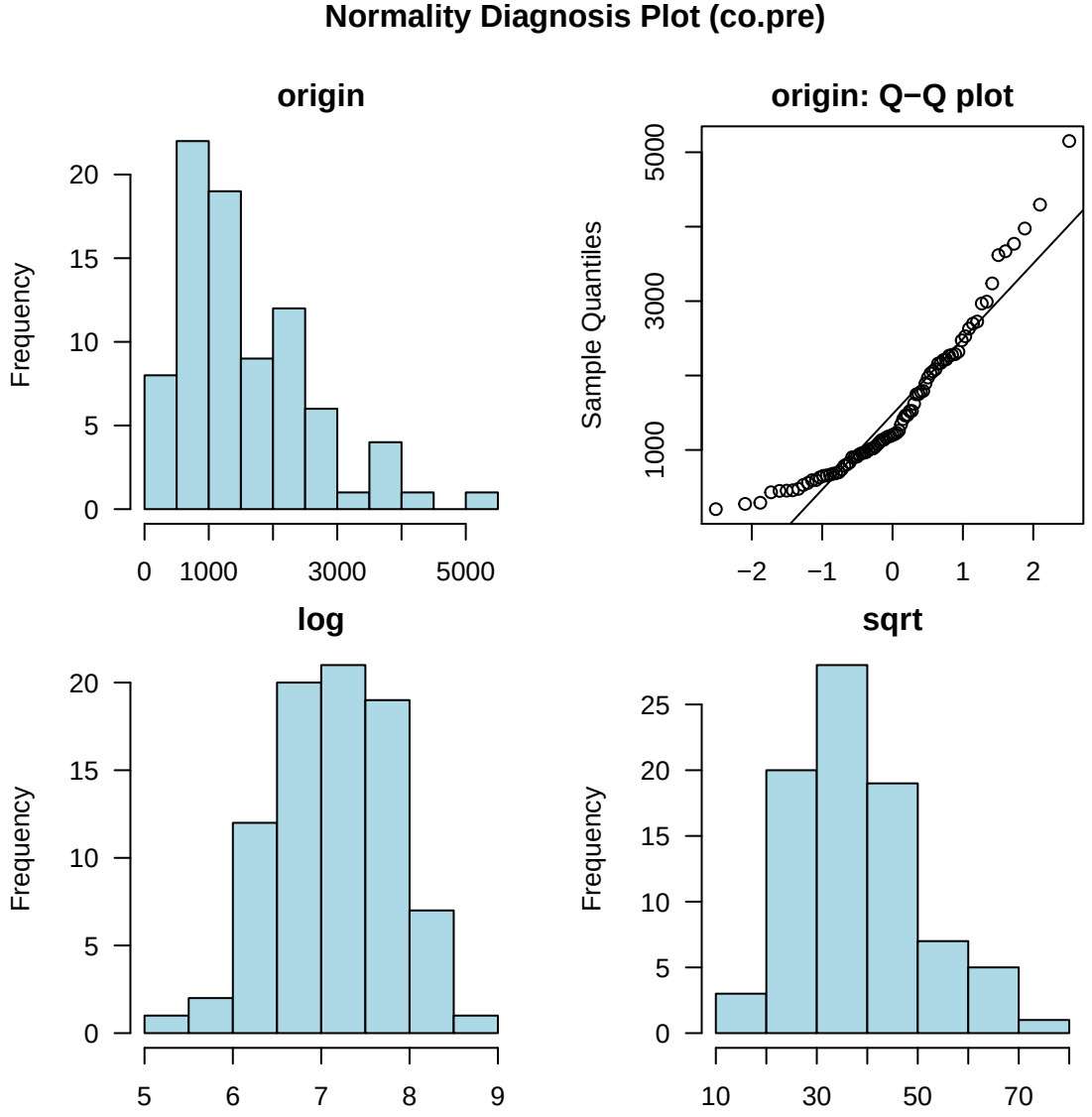
Para analizar el comportamiento general de las variables, es posible observar el valor de *skewness* para la simetría y el valor de *kurtosis* para los valores *outliers* de las variables numéricas como se ha hecho para el biomarcador oxitocina. En este caso, la variable cuyo valor de *skewness* es más alto es *co.pre*, con un valor de 2.08, el doble que el de la variable respuesta. Con el nivel de significancia en 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*.

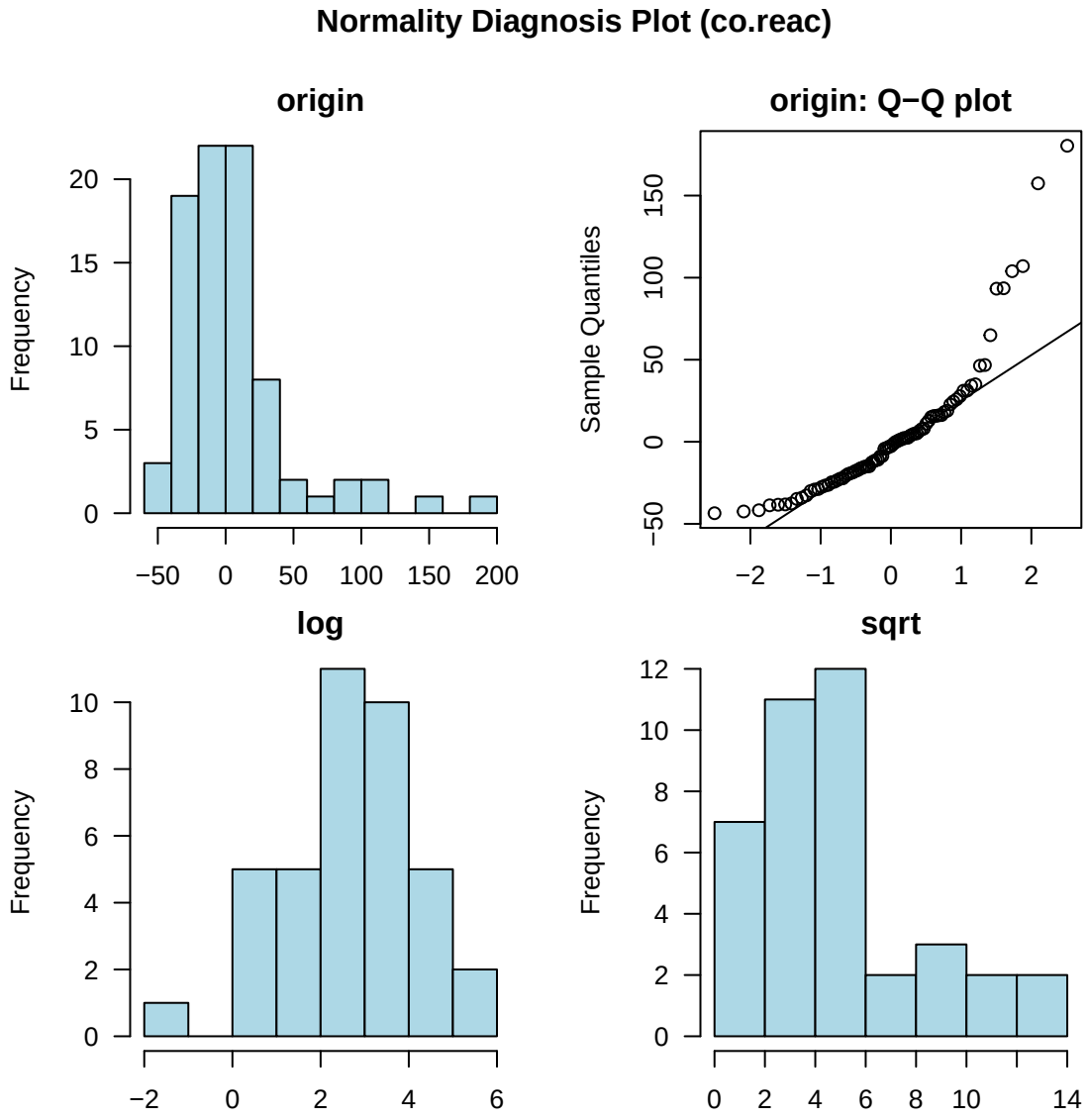
Del test se obtiene que la variable que menos se asemeja a una distribución normal es *med.dos* (p-valor= $1.85 \text{ù} 10^{-12}$), aunque hay que tener en cuenta que muchas de las observaciones de esta variable eran valores NA, y que posterioremente se han convertido a valores nulos (igualados a cero), por lo tanto no es una variable que a simple vista 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*, también con valores imputados para algunas de las observaciones, y tras esta última, está *co.pre*, con un p-valor=$7.27 \text{ù} 10^{-6}$. 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*.

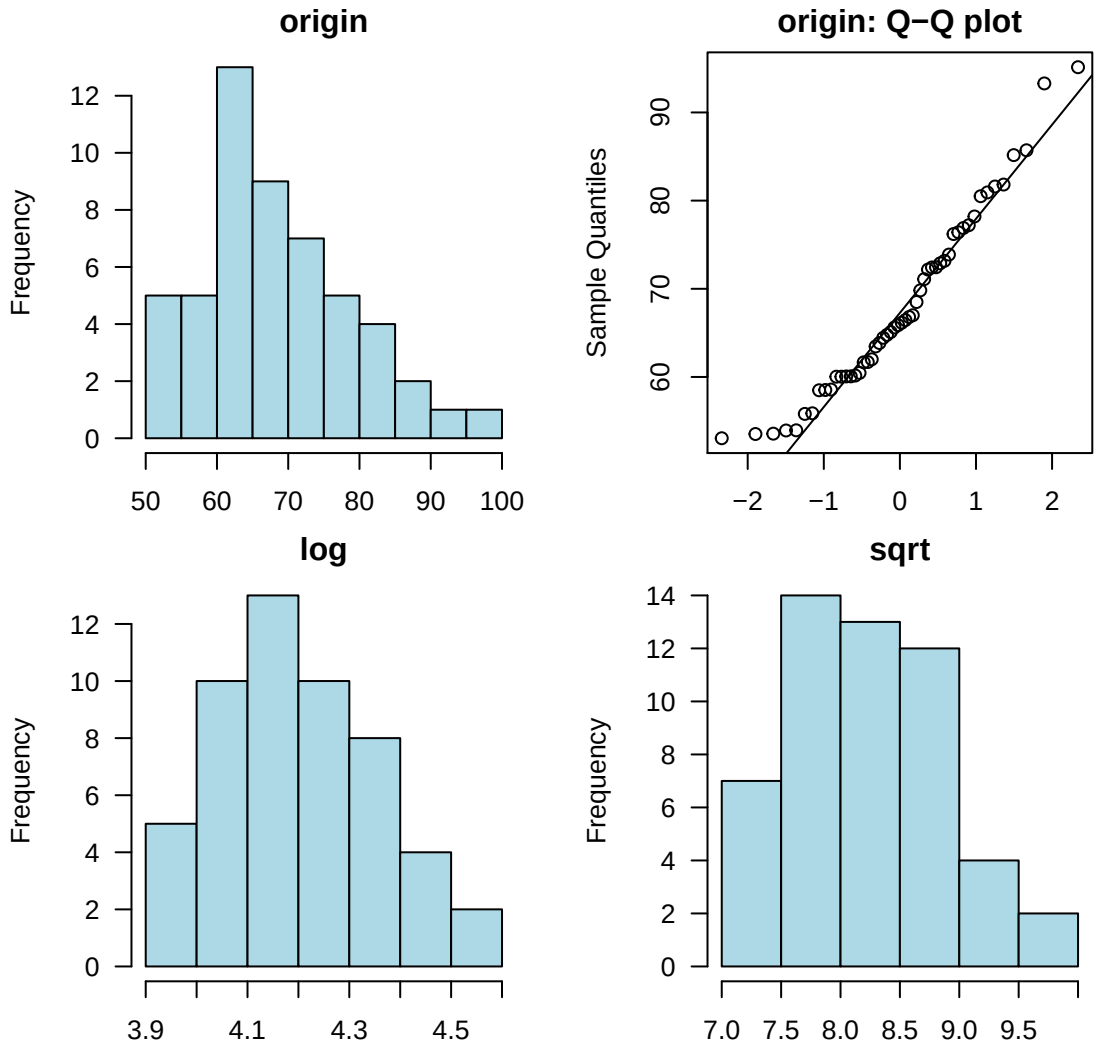
Es aconsejable analizar la distribución de las variables de forma gráfica para ver cómo se comportan y ver las posibles transformaciones para que se asemejen a la distribución normal, y para ello a continuación se muestran unos gráficos obtenidos a partir de la función *plot_normality* para las variables *med.dos*, *co.reac*, *co.post*, *co.pre*, *age*, *hr.bas* y *hr.post*.

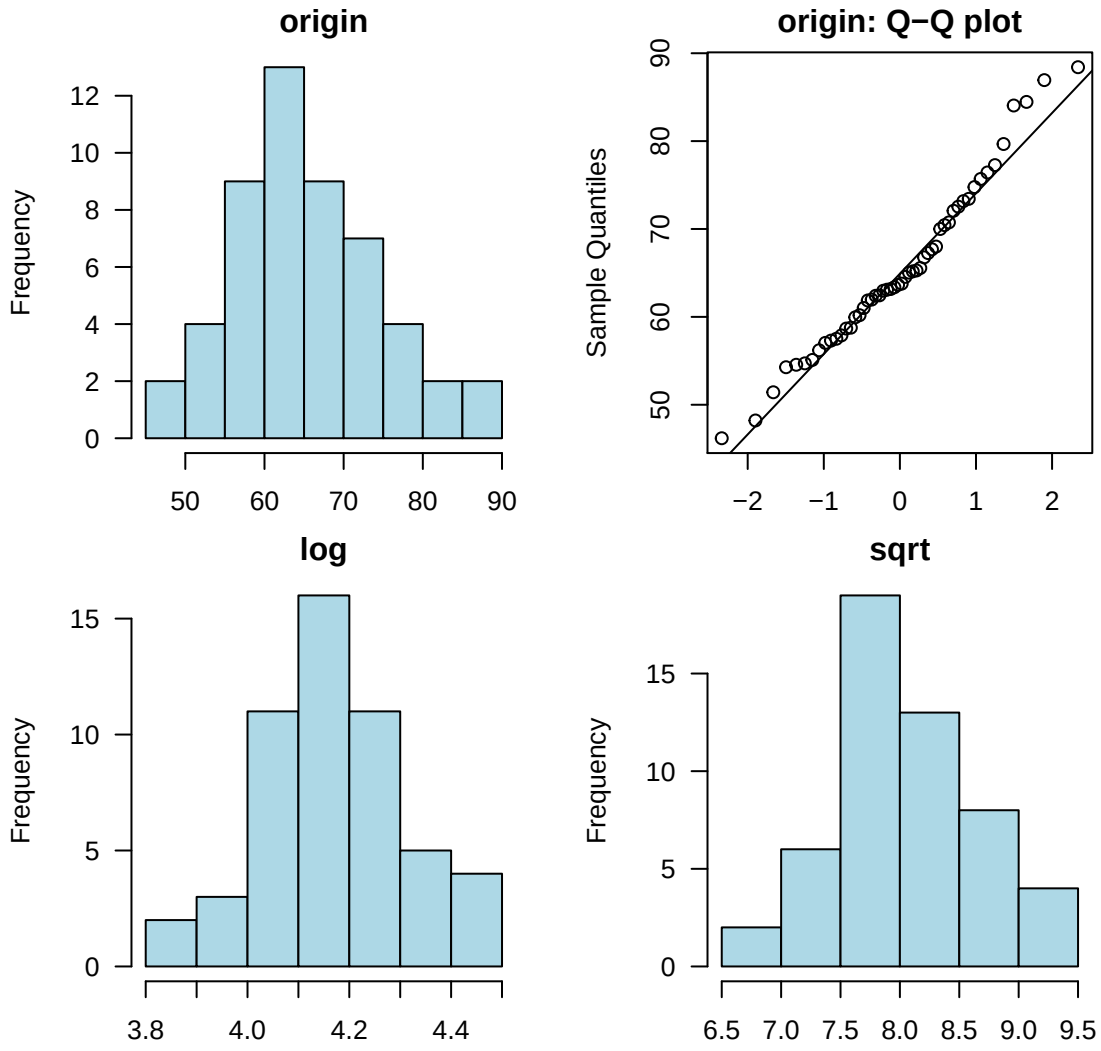



Normality Diagnosis Plot (med.dos)
origin
Frequency
0
10
20
30
40
50
60
0
200
600
1000
origin: Q−Q plot
Sample Quantiles
0
200
600
1000
−2
−1
0
1
2
log
Frequency
0
5
10
15
4
5
6
7
sqrt
Frequency
0
10
20
30
40
50
0
5
10
20
30


Normality Diagnosis Plot (co.pre)
origin
Frequency
0
5
10
15
20
0
1000
3000
5000
origin: Q−Q plot
Sample Quantiles
1000
3000
5000
−2
−1
0
1
2
log
Frequency
0
5
10
15
20
5
6
7
8
9
sqrt
Frequency
0
5
10
15
20
25
10
30
50
70


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


Normality Diagnosis Plot (hr.bas)
origin
Frequency
0
2
4
6
8
10
12
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
12
3.9
4.1
4.3
4.5
sqrt
Frequency
0
2
4
6
8
10
12
14
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 (Figura XXX) confirma que el resultado que se observa gráficamente está relacionado con el p-valor analizado, ya que el histograma que cuya distribución parece 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, es decir, es una mejora considerable). Las variables *hr.bas* y *co.pre* están sesgadas a la derecha sin aplicar ninguna transformación, y si 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 en el test de Shapiro-Wilk. Se observa que en la variable *co.reac* 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 logarímica de esta variable, por lo que sí se podría aceptar con un 5% de significancia la distribución normal de ésta. Tal y como se ha comentado previamente, la variable *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 trasnformarla 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 transfrmación de la raíz cuadrada, no aceptando por lo tanto la distribución normal con un nivel de significancia del 5%.

### 1.1.4 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 ha observado que tiene algunos valores NA, y en algunas variables (ritmos cardiacos sobre todo) el porcentaje de valores faltantes es elevado. Se ha aplicado sobre el conjutno 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 XXX.

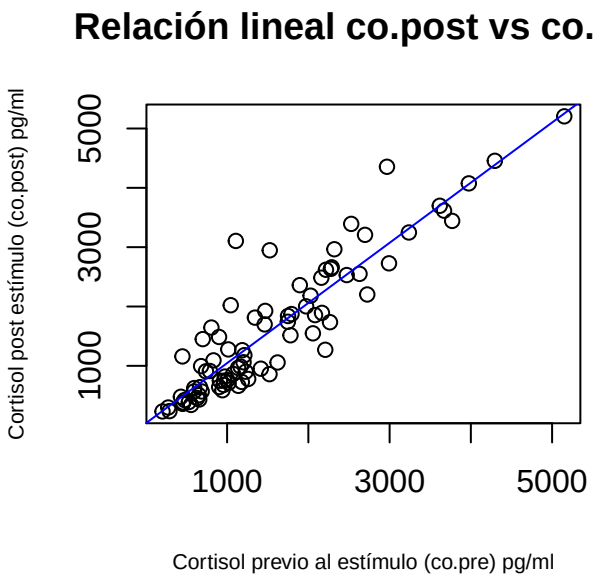
**INCLUIR TABLA DE MATRIZ DE CORRELACIONES**

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 XXX se observa la matriz de correlaciones, y para interpretar si la correlación es fuerte o débil, nos hemos basado en los estudios de Martinez Ortega et al. (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, med.type, med.dos* y *co.meas* muestran una correlación perfecta entre ellas (*coef. = 1*), con la variable *co.pre* una correlación de 0.745 y con la variable respuesta *co.post* un valor similar, 0.774, ya que solo en el estudio de Tas et al. (2018) se utiliza la variable *med.dos* y para las observaciones del otro estudio, éstos valores se han imputado (igualandolos 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 muestra de saliva, otros de sangre), el nivel de correlación entre las tres variables con *co.pre* y *co.post* sería más bajo y se hubiera evitado el patrón que se observa en el análisis.

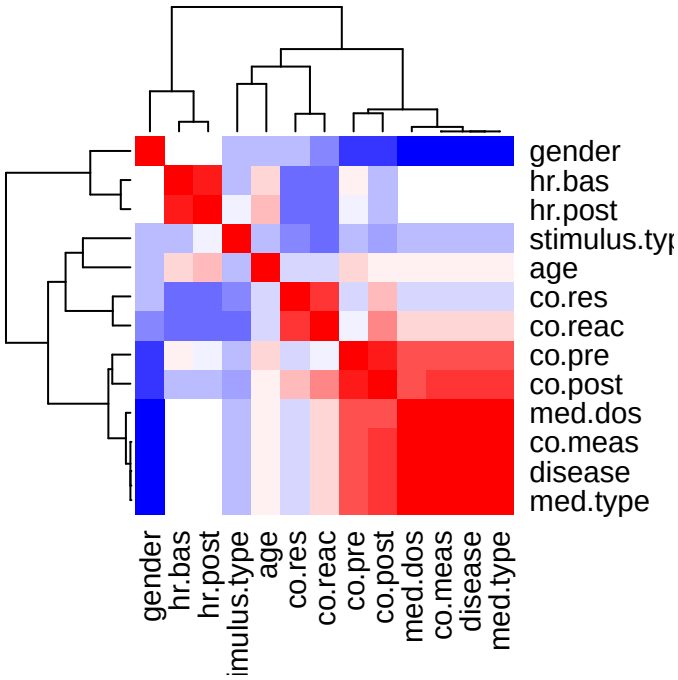
- Ambas variables que miden el ritmo cardiaco (*hr.bas* y *hr.post*) muestran una correlación alta entre ellas, con un valor en el coeficiente de 0.862. Como en el caso del biomarcador oxitocina, se debe eliminar una de ellas a la hora de utilizarlas como variables predictoras.
- Las variables *co.reac* y *co.res* están correlacionadas de forma postiva y además con un valor muy alto (0.785). Es normal ya que *co.res* se genera a partir de los datos obtenidos en *co.reac*.
- La variable *co.pre* y *co.post* están altamente y positivamente correlacionadas entre ellas, con un coeficiente de correlación de 0.885 entre ambas variables. En la figura XXX 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**.




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 variables 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 cuales son significativos al 5%, y en este caso se observa que se obtienen p-valores inferiores a 0.05 en las combinaciones que incluyen a las variables *disease*, *med.type*, *med.dos* y *co.meas* (una vez más por el sesgo de los datosa raíz de los estudios utilizados), y también aquellas que incluyen la variable edad (ya que en uno de los estudios se aplica diferentes estímulos sobre un mismo paciente).

Finalmente, para concluir el análisis de la correlación, a continuación en la Figura XX se muestra un mapa de calor (*heatmap*) donde se puede observar en color rojo las correlaciones más altas entre 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 raiz de los valores obtenidos en *co.reac*) y también en la interacción *co.pre-co.post*. Finalmente, la correlación más significativa se muestra entre las cuatro variables *co.meas*, *disease*, *med.dos* y *med.type*, y con algo menos de intensidad en la interacción entre la variable respuesta y la variable predictora *co.pre*.




### 1.1.5 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 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, ya que en uno de ellos, todos los participantes tienen una enfermedad y toman la misma medicación, y sin embargo, en el otro conjunto de datos, ninguno de los pacientes tiene una enfermedad y tampoco toman ningún medicamento. Ocurre un fenomeno 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 por lo tanto, cada uno de los participantes se repite en el conjunto de datos (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 que se encuentran en el conjunto de datos. 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 menos variables predictoras que las analizadas en 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.

#### 1.1.5.1 Propuesta 1

Para la propuesta 1 por lo tanto 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. 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 pertecer todas ellas a un estudio (y por lo tanto a un tipo de medicicón del cortisol), limita la variabilidad del modelo. Es por ellos por lo que se decide eliminar la variable del modelo aunque su correlación con las otras variables no supongan un problema de independencia. En la siguiente función, se presenta el planteamiento inicial del modelo que mejores resultados ha dado para la predicción del cortisol:

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

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 descretas). La variable respuesta, también se plantea con la transformación logarítmica. En la Figura XXX se muestra el *output* obtenido del sumario del modelo:

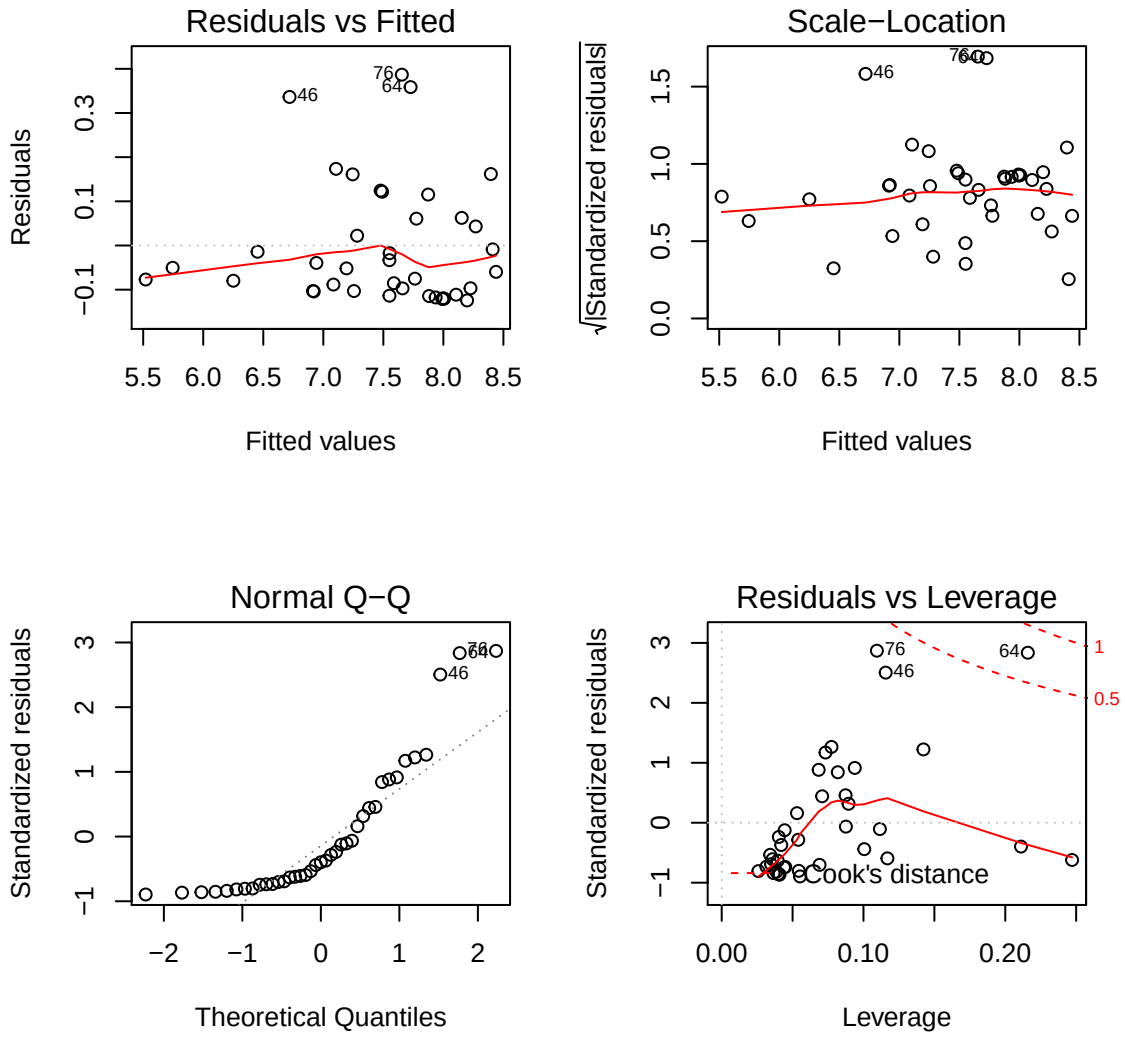
**COPIAR IMAGEN DEL OUTPUT DEL MODELO mod.co.p2**

```
Call:
lm(formula = log(co.post) ~ log(age) + gender + stimulus.type +
    log(co.pre) + log(co.reac), data = data.co)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14920 -0.09494 -0.05338  0.05449  0.35640

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.084939   0.575306   0.148    0.884
log(age)       -0.057796   0.126155  -0.458    0.650
gender2        -0.014180   0.067058  -0.211    0.834
stimulus.type1  0.062917   0.088116   0.714    0.480
stimulus.type2  0.005692   0.077884   0.073    0.942
log(co.pre)     0.986617   0.054268  18.180  < 2e-16 ***
log(co.reac)    0.161720   0.018989   8.517 9.87e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1494 on 32 degrees of freedom
  (45 observations deleted due to missingness)
Multiple R-squared:  0.9624,    Adjusted R-squared:  0.9553
F-statistic: 136.4 on 6 and 32 DF,  p-value: < 2.2e-16

Start:  AIC=-142.02
log(co.post) ~ log(age) + gender + stimulus.type + log(co.pre) +
    log(co.reac)

                Df Sum of Sq    RSS      AIC
- stimulus.type  2    0.0114 0.7253 -145.406
- gender         1    0.0010 0.7149 -143.969
- log(age)       1    0.0047 0.7186 -143.768
<none>                       0.7139 -142.023
- log(co.reac)   1    1.6181 2.3320  -97.856
- log(co.pre)    1    7.3738 8.0877  -49.355
```

```
Step:  AIC=-145.41
log(co.post) ~ log(age) + gender + log(co.pre) + log(co.reac)

                Df Sum of Sq     RSS      AIC
- log(age)       1    0.0011  0.7263 -147.350
- gender         1    0.0074  0.7327 -147.010
<none>                       0.7253 -145.406
+ stimulus.type  2    0.0114  0.7139 -142.023
- log(co.reac)   1    1.6372  2.3625 -101.350
- log(co.pre)    1   14.4380 15.1633  -28.843

Step:  AIC=-147.35
log(co.post) ~ gender + log(co.pre) + log(co.reac)

                Df Sum of Sq     RSS      AIC
- gender         1    0.0078  0.7341 -148.933
<none>                       0.7263 -147.350
+ log(age)       1    0.0011  0.7253 -145.406
+ stimulus.type  2    0.0077  0.7186 -143.768
- log(co.reac)   1    1.6386  2.3649 -103.310
- log(co.pre)    1   15.6144 16.3407  -27.926

Step:  AIC=-148.93
log(co.post) ~ log(co.pre) + log(co.reac)

                Df Sum of Sq     RSS      AIC
<none>                       0.7341 -148.933
+ gender         1    0.0078  0.7263 -147.350
+ log(age)       1    0.0014  0.7327 -147.010
+ stimulus.type  2    0.0133  0.7208 -145.648
- log(co.reac)   1    1.6329  2.3670 -105.275
- log(co.pre)    1   17.8383 18.5724  -24.933


Call:
lm(formula = log(co.post) ~ log(co.pre) + log(co.reac), data = data.co)

Coefficients:
 (Intercept)   log(co.pre)  log(co.reac)
     -0.2872        1.0147        0.1595


Call:
lm(formula = log(co.post) ~ log(co.pre) + log(co.reac), data = data.co)

Residuals:
     Min       1Q   Median       3Q      Max
-0.12433 -0.10308 -0.05186  0.06161  0.38673

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -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

Residual standard error: 0.1428 on 36 degrees of freedom
  (45 observations deleted due to missingness)
Multiple R-squared:  0.9613,    Adjusted R-squared:  0.9592
F-statistic: 447.3 on 2 and 36 DF,  p-value: < 2.2e-16
```

En la imagen anterior (Figura XXX) se puede observar que finalmente, las unicas 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 ambas variables mencionadas. El valor del $R^2$ es 0.9592, siendo 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. A continuación, en la figura XX, se muestran cuatro gráficos diferentes que describen los residuos del modelo *mod.co.p2*.



En la figura XXX se observa mediante el gráfico de *Scale Location* que el modelo parece que si que 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 linearidad y normalidad, aunque se ha comprobado que de eliminarlos, sí que se conseguiría un valor más alto respecto al $R^2$, pero que no mejoraría las suposiciones de linearidad ni normalidad gráficamente ni en los test. Por lo tanto, no se considera que eliminar los valores influyentes (en concreto las observaciones 46, 64 y 76) del conjunto de datos sea efectivo en este caso.

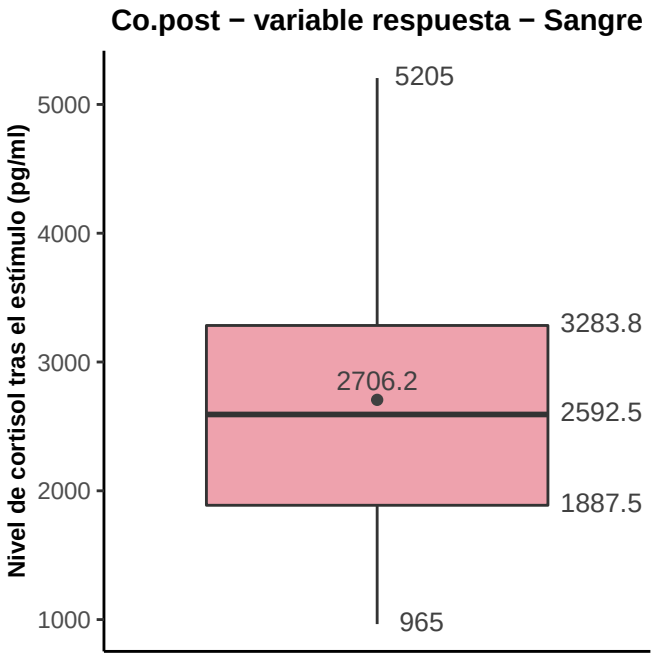
Al aplicar el test de Sahpiro-wilk en los residuos del modelo, donde se quiere verificar si éstos siguen o no una distribución normal, se observa que el p-valor obtenido tiene un valor de $1.26 \mathring{u} 10^{-5}$, por lo tanto se rechaza la hipótesis nula del test y no se asume la normalidad de los residuos. El no cumplir con la suposición de la normalidad ni de la linearidad (mostrada en la Figura XXX), es suficiente para rechazar este modelo para predecir el valor del *co.post* con el conjunto de datos general del cortisol. Ninguna de las transformaciones de los datos que se ha llevado a cabo ha cumplido con la hipótesis de la normalidad, y cabe destacar que el modelo *mod.co.p2* es el que ha mostrado mejores resultados en general, obteniendo valores AIC y BIC mucho más bajos que para los demás modelos (explicados en el Anexo C del documento). 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 de recoger la muestra, en sangre o en saliva, tal y como se explica en los siguientes subapartados.

#### 1.1.5.2 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.

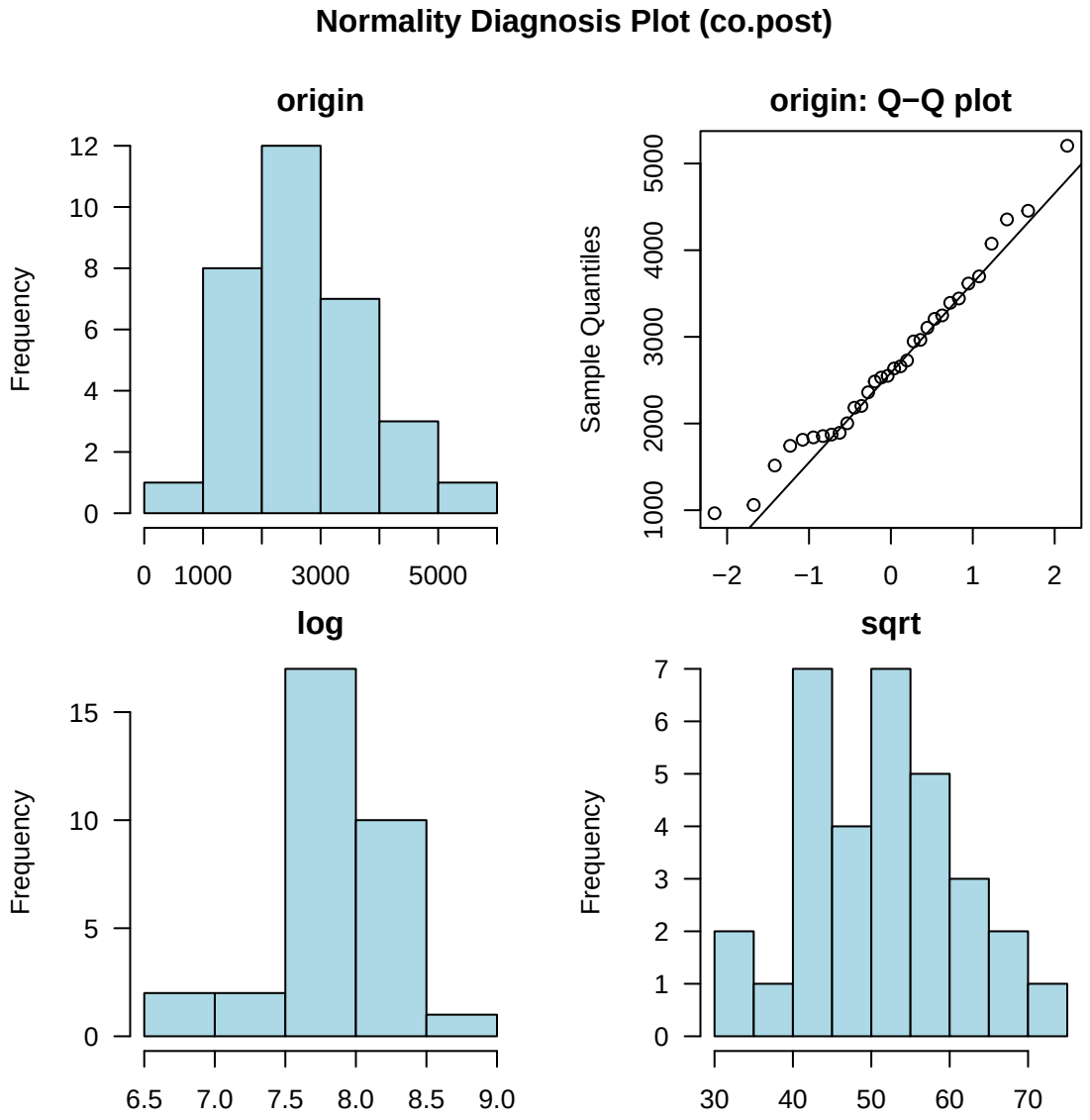
##### 1.1.5.2.1 Sangre

Con el objetivo de generar el modelo utilizando únicamente observaciones de la sangre, del mismo modo que para la saliva, se ha separado el conjunto de datos original *data.co* según los valores en la variable *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 el estudio de donde se han obtenido las observaiones en sangre no se ha medido el ritmo cardiaco 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 XXX) como numéricamente en la tabla XXX. 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.



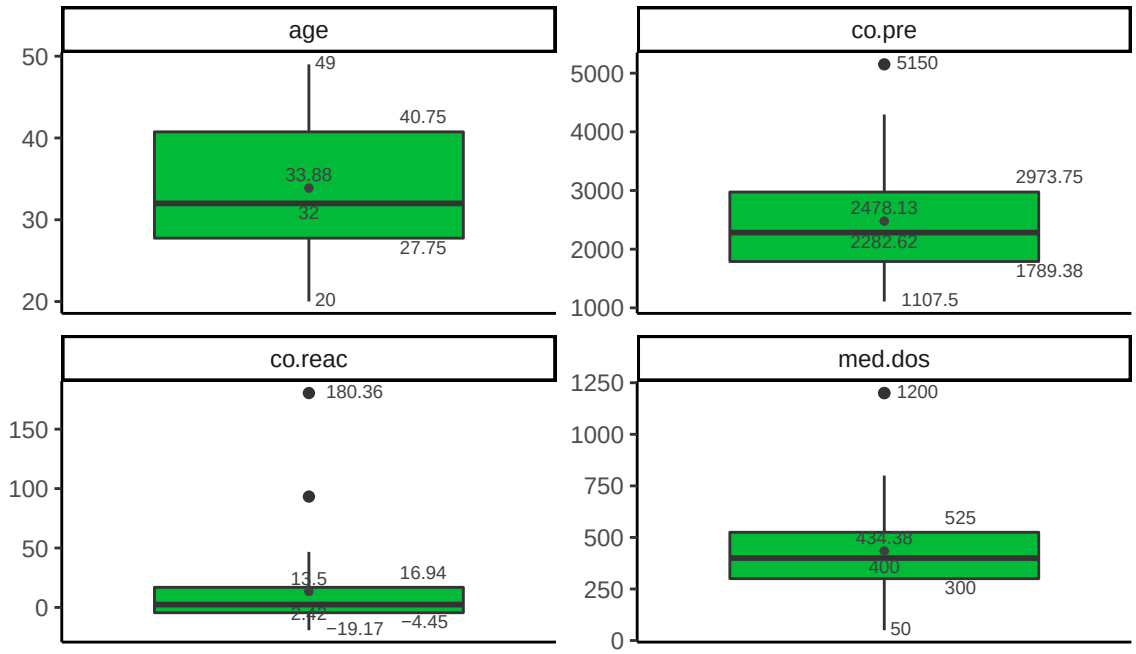

**METER TABLA DE WORD AQUI**

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 *co.post*. En la figura XXX 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 los datos se refiere. Además, los puntos del gráfico Q-Q parece que en general están sobrepuestos en la línea de la normal, aunque en la cola haya unos puntos que difieren.

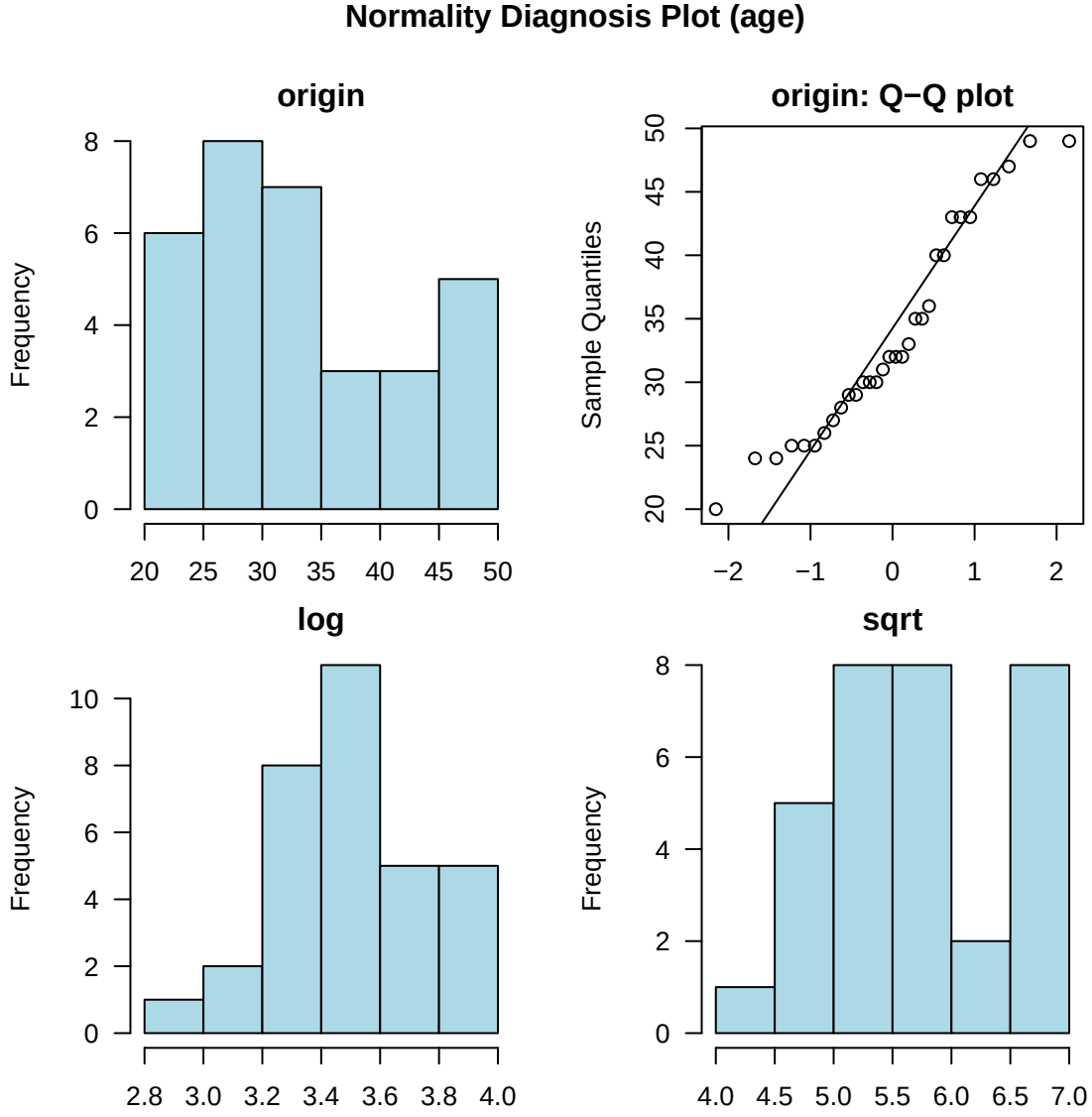


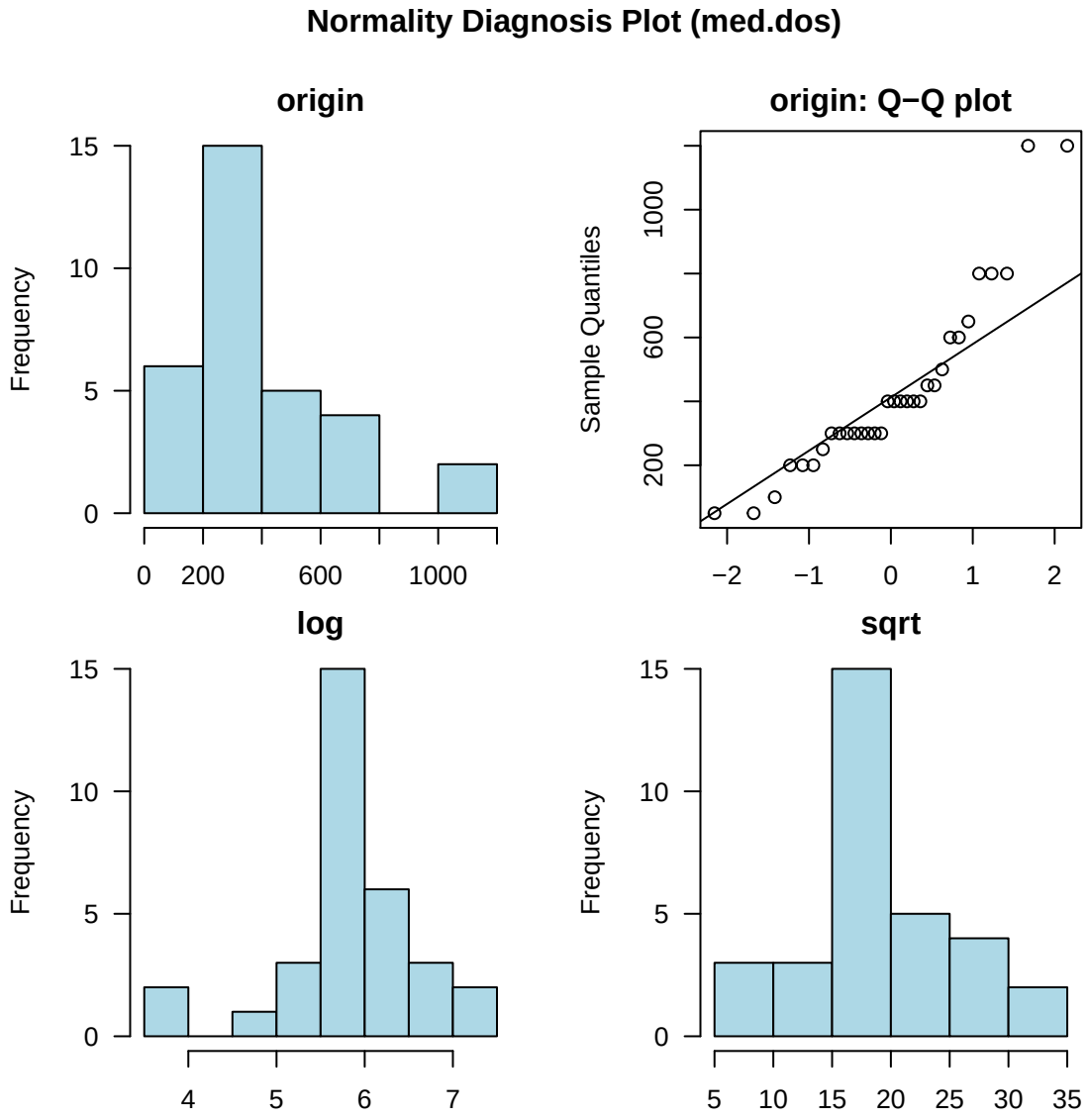

Respecto a las variables predictoras, en la siguiente figura XXX se muestra la distribución de las mismas, y en la tabla XXX se resumen los datos más significativos de cada una de las variables para este conjunto de datos, aunque estos datos ya se han mostrado por grupos en las tablas XXX y xxx (para la variable edad).

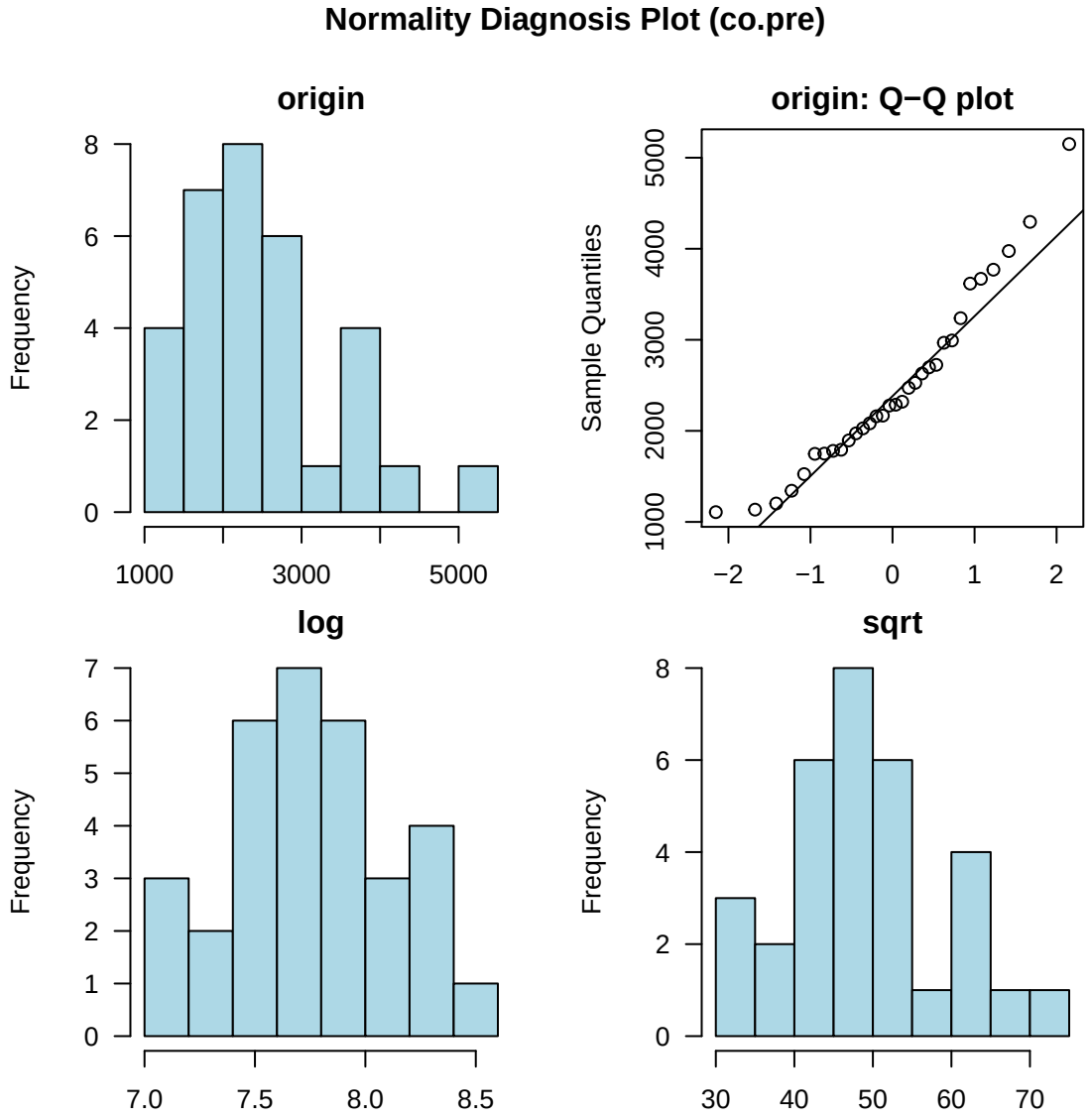
**METER TABLA**

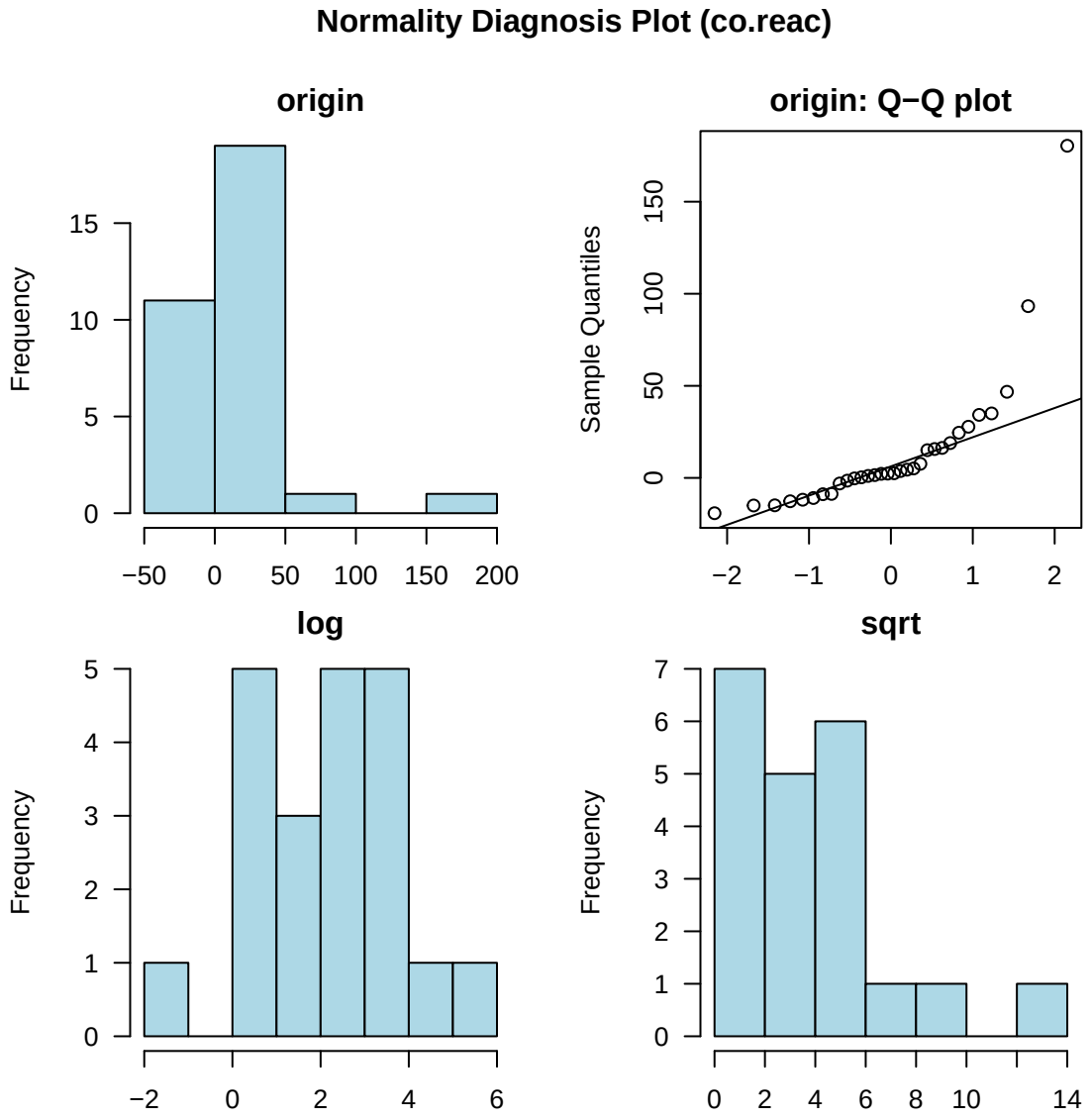



Respecto a la distribución normal de estas variables, sin aplicar ninguna transformación sobre ellas, la única variable significativa al 5% es *co.pre*, con un p-valor ligeramente superior al 5% (p-valor=0.083) y por lo tanto se aceptaría la distribución normal sobre ella. En la figura XXX se observa 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*, con un p-valor=0.01. La distribución de estas variables se muestra en los siguientes gráficos:


Normality Diagnosis Plot (age)
origin
Frequency
origin: Q−Q plot
Sample Quantiles
log
Frequency
sqrt
Frequency

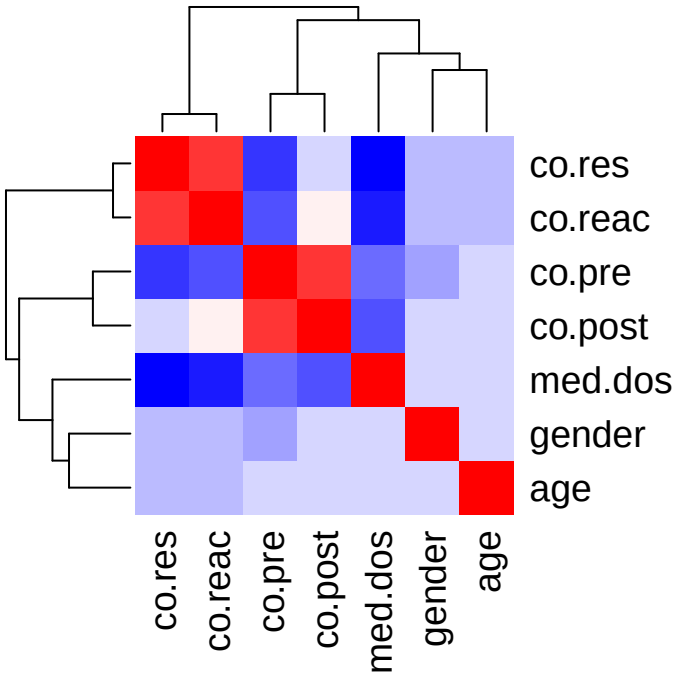

Normality Diagnosis Plot (med.dos)
origin
Frequency
0
5
10
15
0
200
600
1000
origin: Q−Q plot
Sample Quantiles
200
600
1000
−2
−1
0
1
2
log
Frequency
0
5
10
15
4
5
6
7
sqrt
Frequency
0
5
10
15
5
10
15
20
25
30
35


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




Finalmente, respecto a las correlaciones entre las variables, en la tabla XX se muestran los valores de los coeficientes de correlación para los datos del conjunto de datos único 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 XXX.

**INCLUIR TABLA DE CORRELACIÓN DE WORD AQUÍ**




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 son *co.res* y *co.reac* para el conjunto de datos de la sangre y 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 el desarrollo de este apartado, se han planteado cuatro modelos diferentes, y en la presente sección se muestra el modelo con mejores resultados para predecir la variable respuesta, y que además cumple con los supuestos para un modelo lineal. Los otros tres modelos planteados, se muestran en el anexo XXX de este documento. El modelo que se plantea se denomina *mod.co.sngr3* y la fórmula que se ha planteado es la siguiente:

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

El modelo está compuesto por las variables predictoras *co.pre*, *age*, *co.reac* y *med.dos*, y la variable respuesta (*co.post*) transformada logarítmicamente. En un primer planteamiento, se había incluido la variable predictora *gender*, pero no era significativa, y por lo tanto, tras aplicar la función *stepAIC* para llevar a cabo la selección de los predictores del modelo, se ha decidido eliminarla, ya que no era significativa y por lo tanto no tenía un efecto sobre la variable respuesta *co.post*. En la figura XXX se muestra el *output* obtenido del modelo:

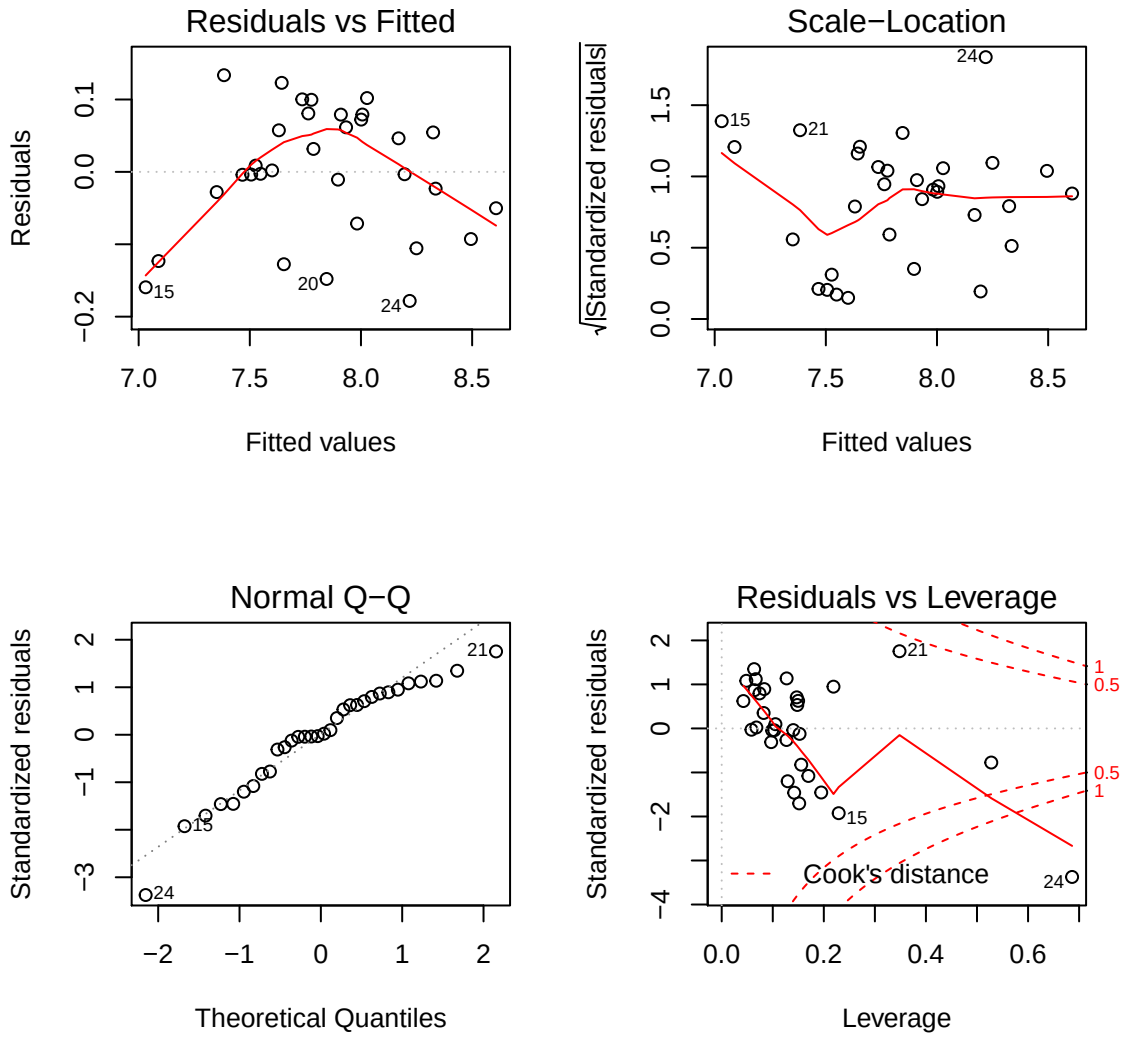
**COPIAR IMAGEN DEL OUTPUT DEL MODELO mod.co.sngr3**

```
Call:
lm(formula = log(co.post) ~ co.pre + age + gender + co.reac +
    med.dos, data = data.co.sngr)

Residuals:
      Min        1Q    Median        3Q       Max
-0.179564 -0.052169 -0.000536  0.072542  0.131841
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.746e+00  8.606e-02  78.383  < 2e-16 ***
co.pre       3.849e-04  1.882e-05  20.452  < 2e-16 ***
age          4.845e-03  2.146e-03   2.257 0.032621 *
gender2      7.218e-03  3.564e-02   0.203 0.841072
co.reac      5.370e-03  4.980e-04  10.784  4.3e-11 ***
med.dos     -2.428e-04  6.431e-05  -3.776 0.000837 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09606 on 26 degrees of freedom
Multiple R-squared:  0.949, Adjusted R-squared:  0.9392
F-statistic: 96.81 on 5 and 26 DF,  p-value: 5.902e-16

Start:  AIC=-144.58
log(co.post) ~ co.pre + age + gender + co.reac + med.dos

          Df Sum of Sq    RSS      AIC
- gender   1    0.0004 0.2403 -146.534
<none>                 0.2399 -144.584
- age      1    0.0470 0.2869 -140.857
- med.dos  1    0.1315 0.3714 -132.594
- co.reac  1    1.0731 1.3130  -92.190
- co.pre   1    3.8597 4.0996  -55.755

Step:  AIC=-146.53
log(co.post) ~ co.pre + age + co.reac + med.dos

          Df Sum of Sq    RSS      AIC
<none>                 0.2403 -146.534
+ gender   1    0.0004 0.2399 -144.584
- age      1    0.0504 0.2907 -142.444
- med.dos  1    0.1324 0.3727 -134.488
- co.reac  1    1.1141 1.3544  -93.196
- co.pre   1    3.8888 4.1291  -57.526


Call:
lm(formula = log(co.post) ~ co.pre + age + co.reac + med.dos,
    data = data.co.sngr)

Coefficients:
(Intercept)       co.pre          age      co.reac      med.dos
  6.7446381    0.0003852    0.0049262    0.0053879   -0.0002409


Call:
lm(formula = log(co.post) ~ co.pre + age + co.reac + med.dos,
    data = data.co.sngr)

Residuals:
      Min        1Q    Median        3Q       Max
-0.178289 -0.055495 -0.000339  0.073959  0.133559
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.745e+00  8.432e-02  79.985  < 2e-16 ***
co.pre       3.852e-04  1.843e-05  20.904  < 2e-16 ***
age          4.926e-03  2.071e-03   2.379 0.024691 *
co.reac      5.388e-03  4.816e-04  11.189 1.21e-11 ***
med.dos     -2.408e-04  6.244e-05  -3.857 0.000645 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09434 on 27 degrees of freedom
Multiple R-squared:  0.9489,    Adjusted R-squared:  0.9414
F-statistic: 125.5 on 4 and 27 DF,  p-value: < 2.2e-16
```

En la figura xxx se observa que el valor de $R^2$ ajustado es 0.94, 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 XXX, se muestran cuatro gráficos diferentes que describen los residuos del modelo *mod.co.sngr3*.




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. En este caso, no parece que los residuos se distribuyan alreadedor de la línea horizontal de manera homogenea, 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 linearidad 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 si 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 homocedasticos, 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. En este caso, se observa que hay algunos valores outliers, y que además uno de ellos está separado por la distancia de Cook = 1.

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 en la misma figura. 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.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 analizado es constante (homocedastico).

- **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. 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 preditoras del modelo son cercanos a uno, esto es suficiente para rechazar el principio de multicolinealidad en los residuos del modelo analizado.

### 1.1.6 Conclusión modelo y comparación

El modelo *mod.co.sngr3* es él único modelo planteado para el cortisol (utilizando la base de datos de la sangre) que cumple con los supuestos para un modelo lineal, aunque la suposición de linearidad observada en el gráfico de los residuos no sea óptima a simple vista y tampoco sea el modelo cuyo valor de $R^2$ ajustado sea más elevado. Es también el modelo que más variables predictoras significativas tiene en comparación con los planteados en el Anexo XXX. Para comprobar que efectivamente es el modelo con mejores resultados para predecir el nivel de *co.post*, se han aplicado los métodos AIC y BIC para la comparación entre modelos, y entre todas las combinaciones posibles, es con el que se han obtenido valores más bajos, lo cual es deseable a

la hora de aplicar las funciones mencionadas. Finalmente, la ecuación del modelo *mod.co.sngr3* obtenida es la siguiente:

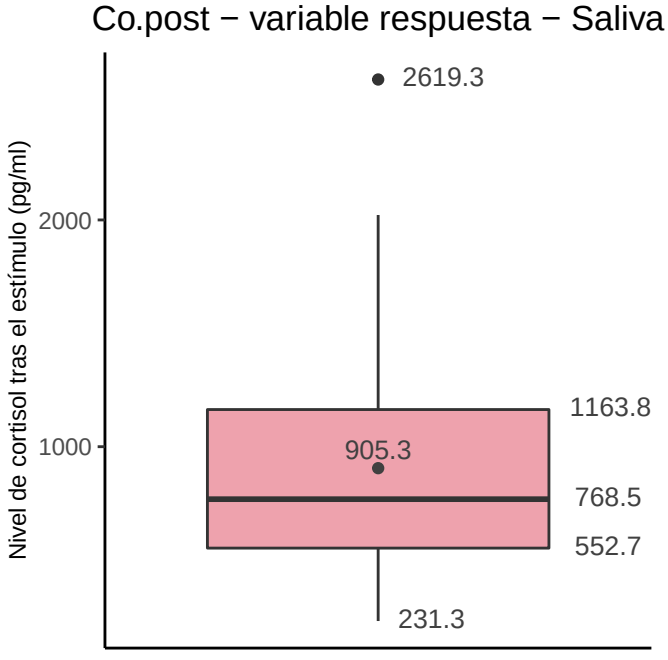
$$log(Y) = 6.745 + 0.00039\ X_1 + 0.00493\ X_2 + 0.00539\ X_3 - 0.00024\ X_4 + \epsilon$$

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*.

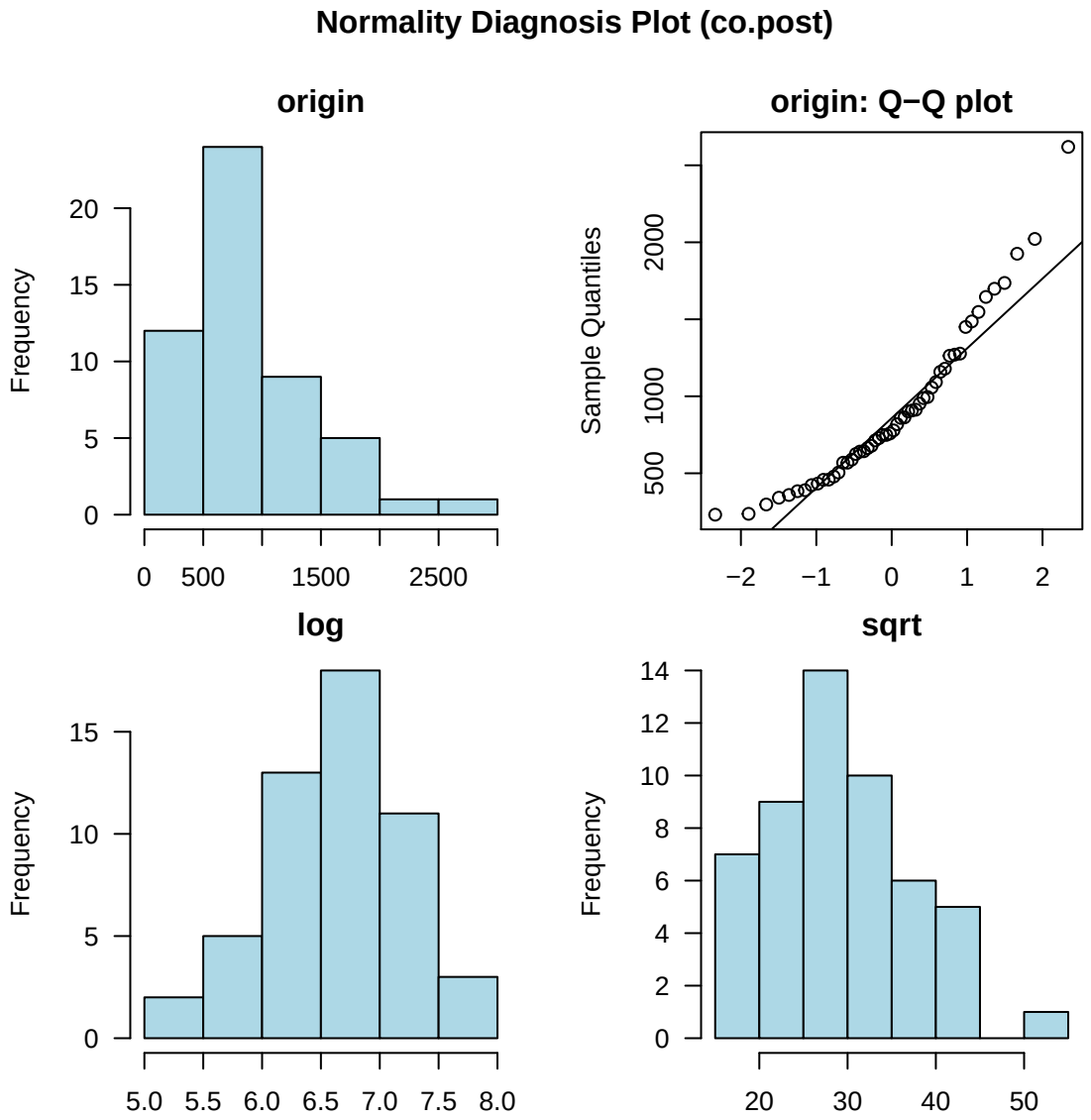
#### 1.1.6.0.1 Saliva

Para generar un modelo útilizando 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* 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 raiz 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, dado que el número de observaciones ha disminuido, a continuación se vuelven a mostrar estas variables. Finalmente, también se volverá a analizar la correlación entre variables, ya que en este caso la reducción de la base de datos sí que modificará los coeficientes de correlación entre las variables.



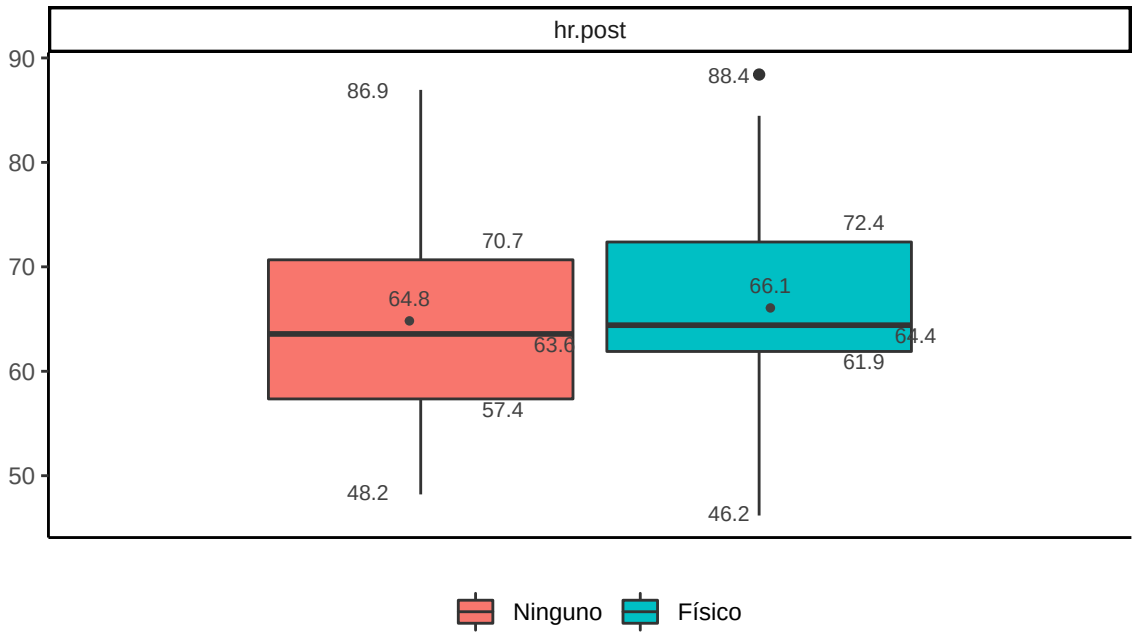
**METER TABLA DE WORD AQUI**

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, ésto se analiza en la imagen XXX, 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.

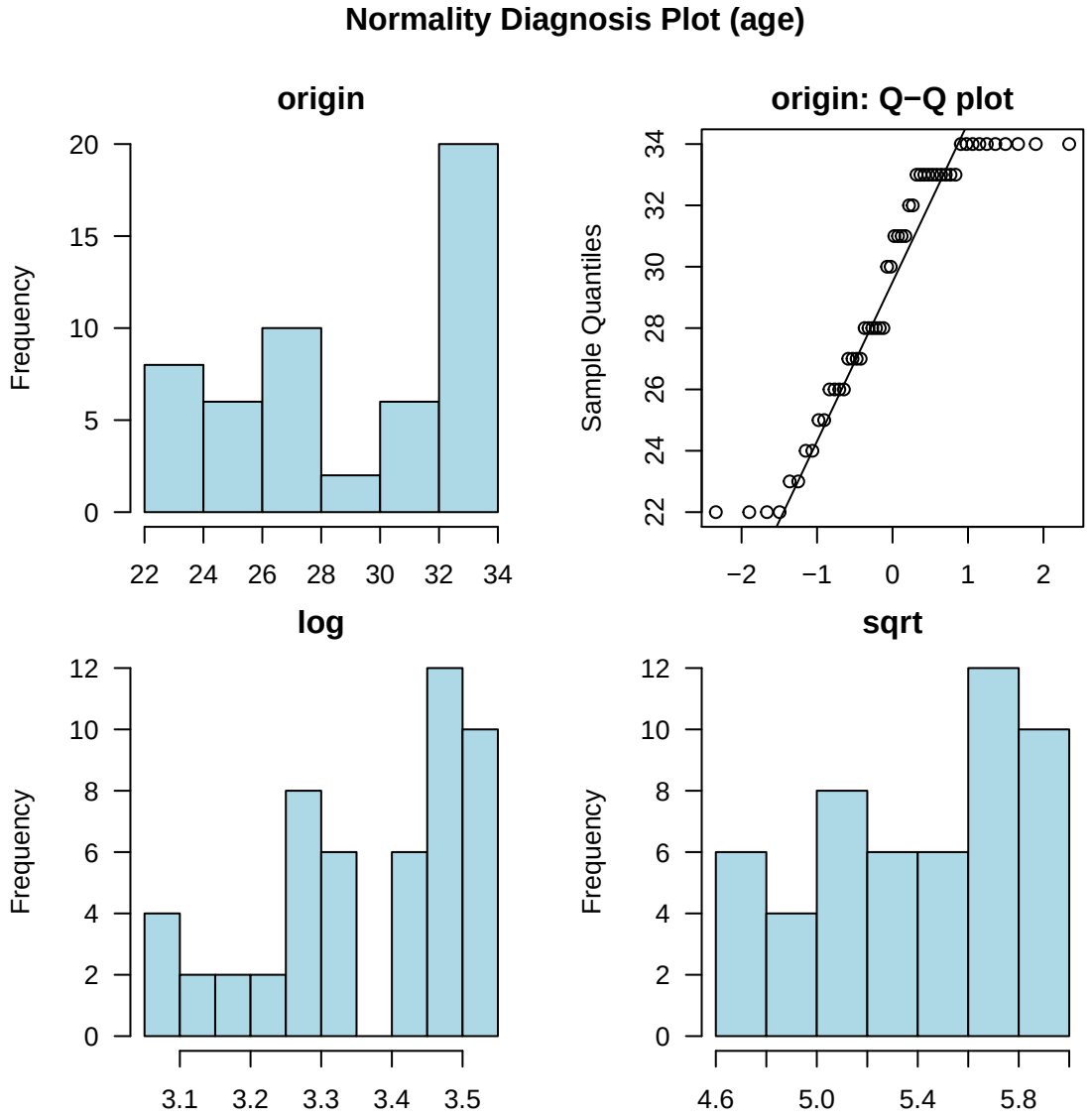


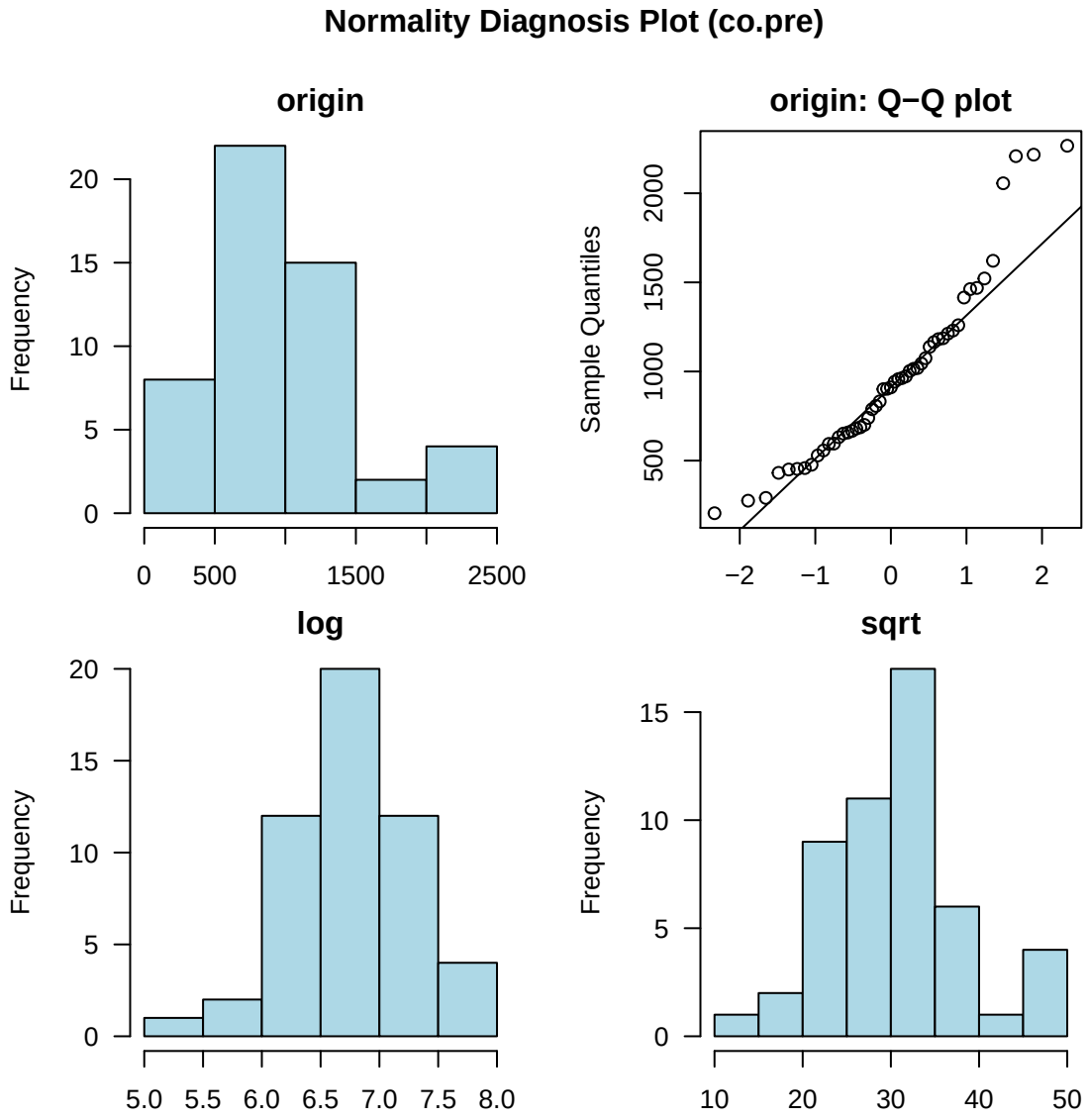

Respecto a las variables predictoras, en la siguiente figura XXX se muestra la distribución de las mismas, y en la tabla XXX se resumen los datos más significativos de cada una de las variables para este conjunto de datos. Los datos se muestran de manera general, puesto que en la tabla xxx mostrada anteriormente ya se ha especificado el EDA para cada uno de los grupos del conjunto general.

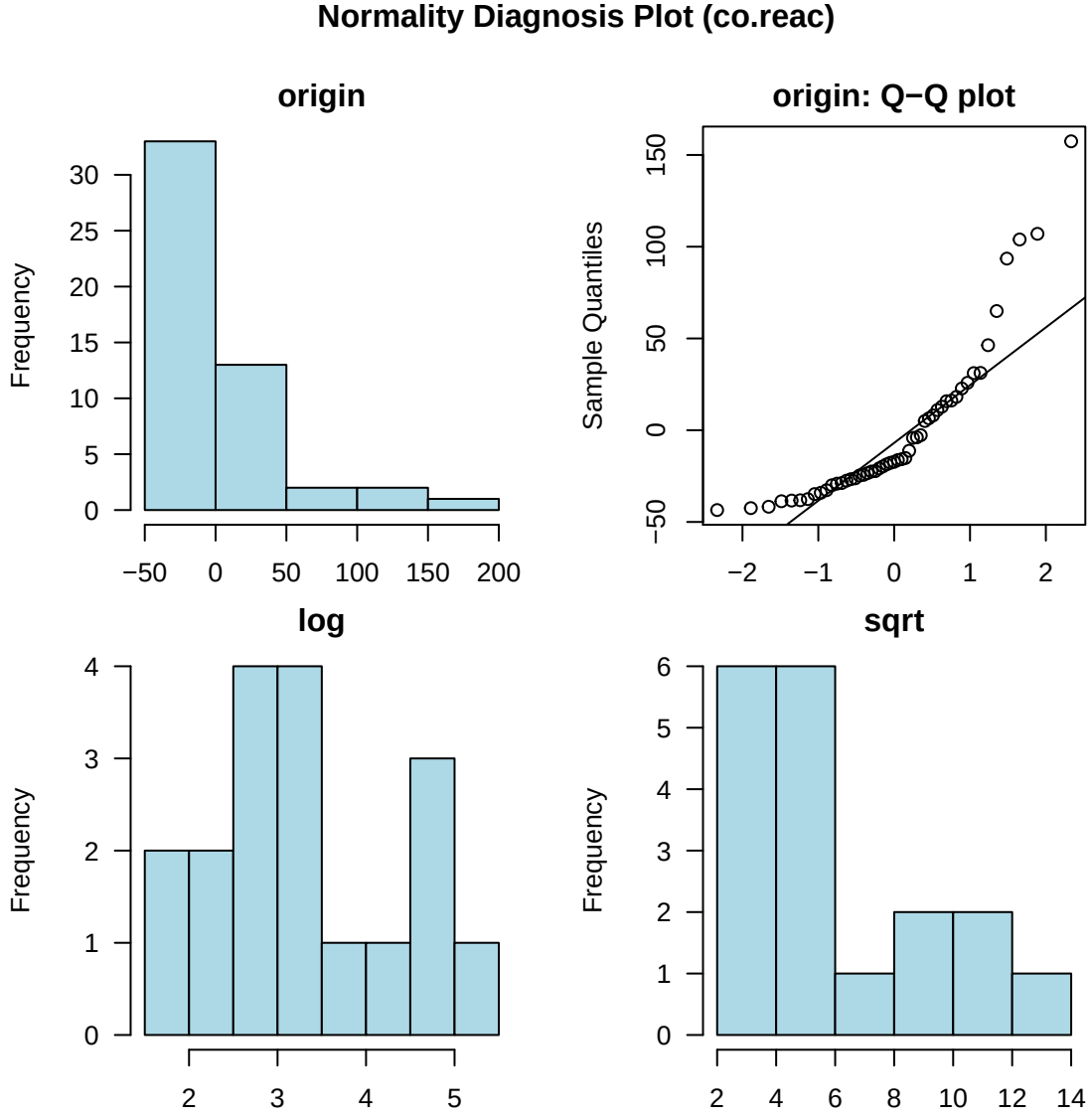
**METER TABLA**

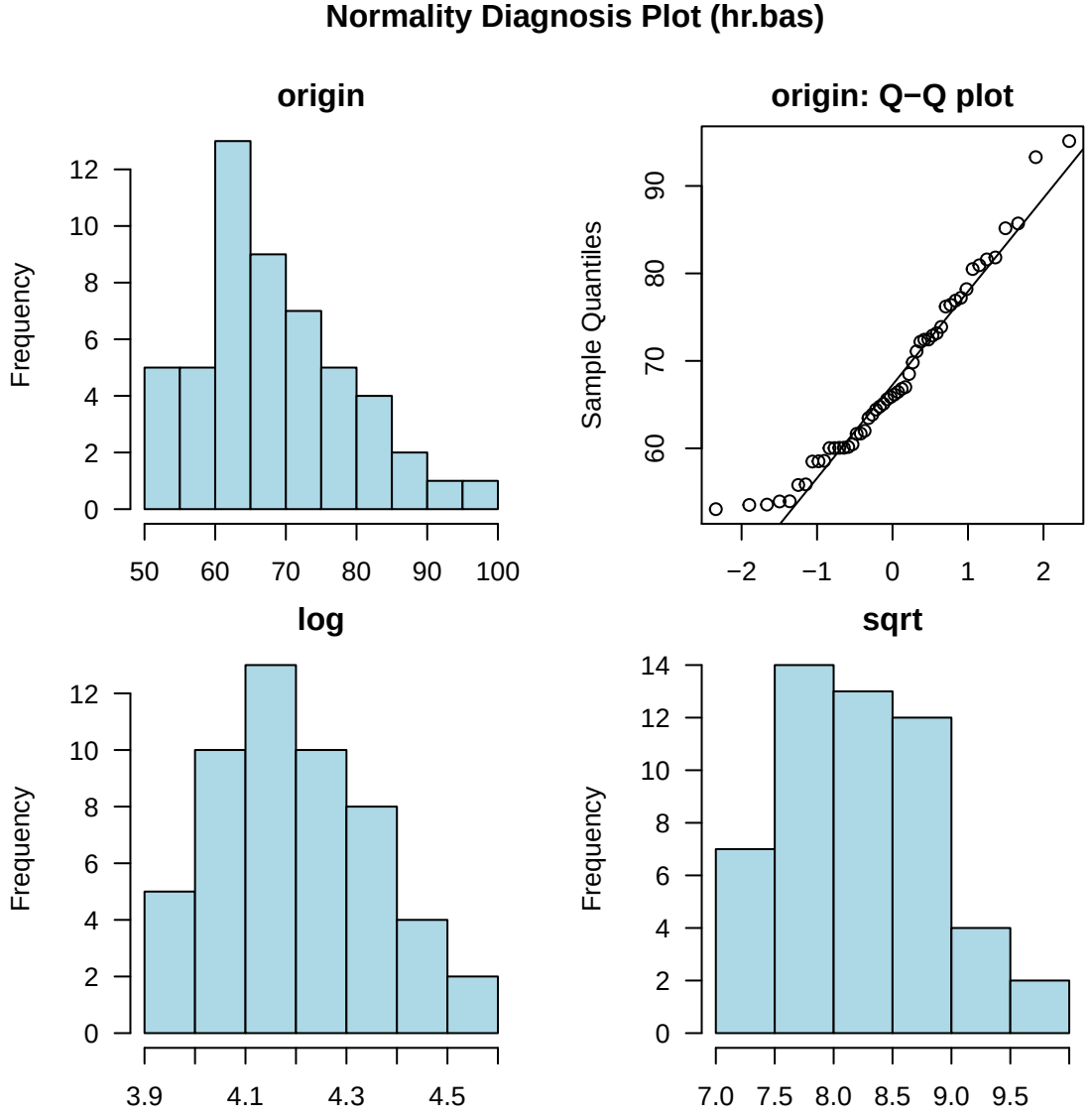


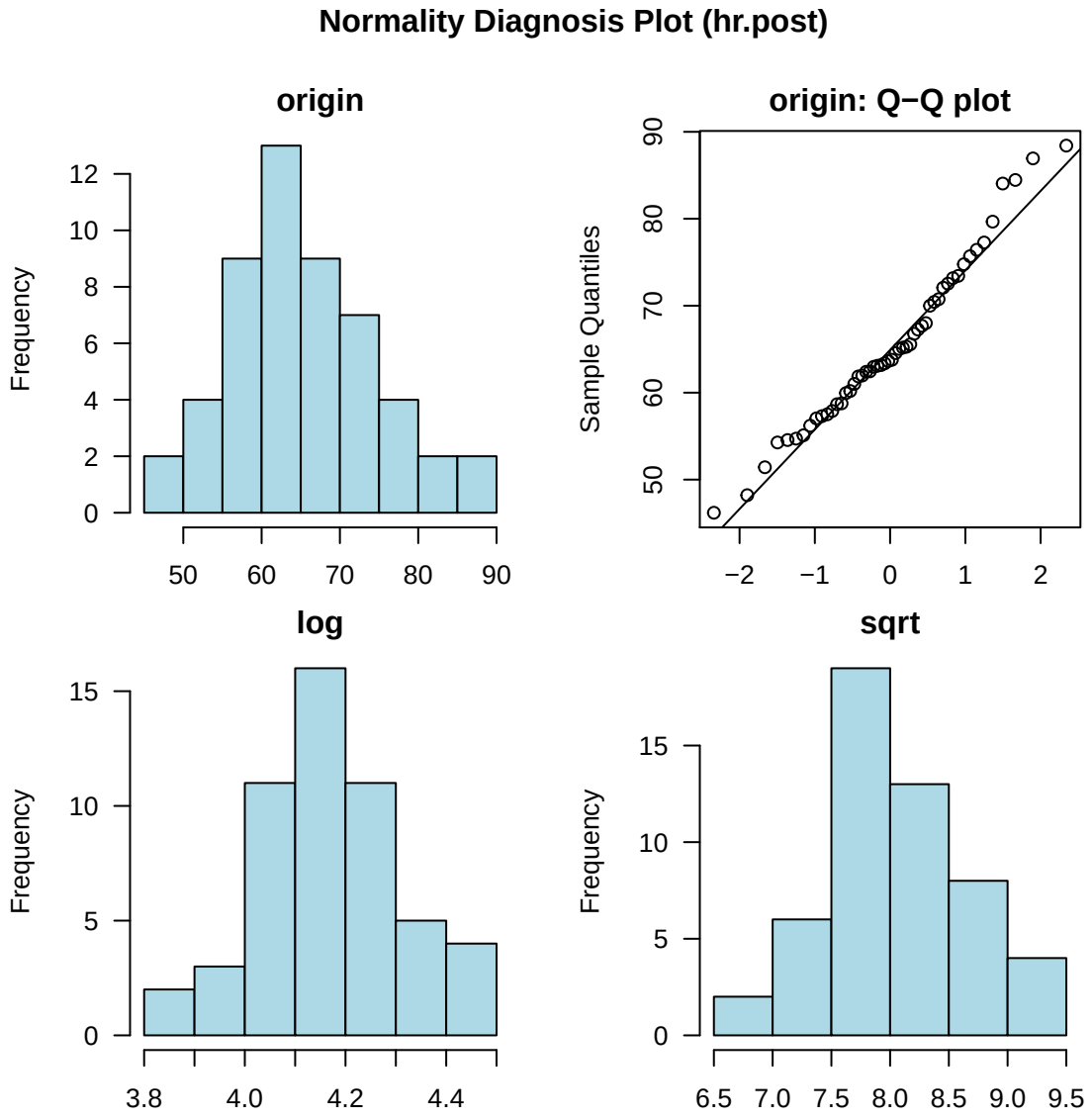

La distribución de las variables *hr.post* y *hr.bas* es las mismas en este conjunto de datos que en el conjunto de datos para el cortisol general, puesto que únicamente teníamos observaciones de estas variables en las muestras obtenidas mediante la saliva. La distribución de las variables *co.reac*, *age*, y *co.pre* ha variado respecto al conjunto de datos original, 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 si 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.


Normality Diagnosis Plot (age)
origin
Frequency
origin: Q−Q plot
Sample Quantiles
log
sqrt

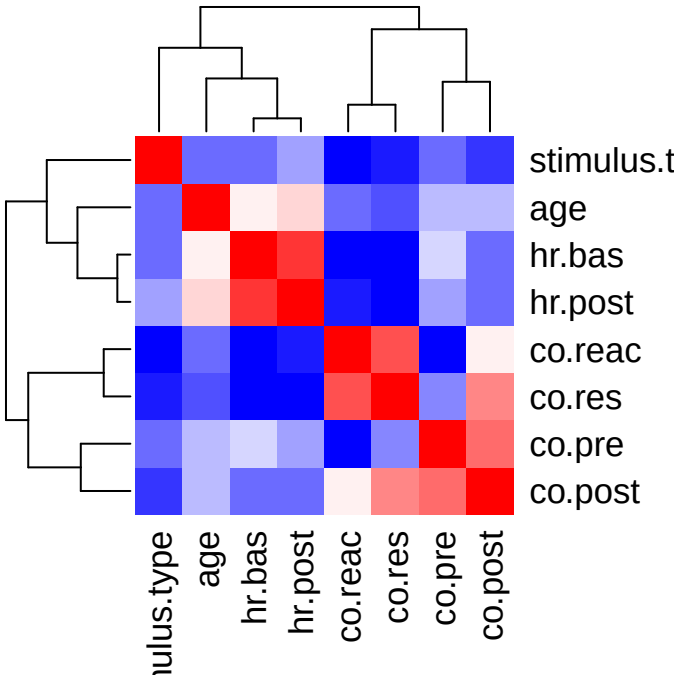

Normality Diagnosis Plot (co.pre)
origin
Frequency
0
5
10
15
20
0
500
1500
2500
origin: Q−Q plot
Sample Quantiles
500
1000
1500
2000
−2
−1
0
1
2
log
Frequency
0
5
10
15
20
5.0
5.5
6.0
6.5
7.0
7.5
8.0
sqrt
Frequency
0
5
10
15
10
20
30
40
50


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


Normality Diagnosis Plot (hr.bas)
origin
Frequency
0
2
4
6
8
10
12
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
12
3.9
4.1
4.3
4.5
sqrt
Frequency
0
2
4
6
8
10
12
14
7.0
7.5
8.0
8.5
9.0
9.5




Finalmente, respecto a las correlaciones entre las variables, a continuación se muestra el mapa de calor (*heatmap*) obtenido a partir del conjunto de datos. En la tabla XX 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). 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.




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 y en la figura XXX anterior se ha mostrado que las variables 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 inclumpliría la condición de independencia entre ellas. Por lo tanto, en el caso de la pareja *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* (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.

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 logaritmicas. 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 XXX del documento), es con el que se han obtenido valores más bajos para las funciones de AIC y BIC. En el planteamiento inicial, el modelo estaba compuesto por las variables numéricas *co.pre*, *age*, *co.reac* y *hr.post* (transformadas logarítmicamente) y la variable predictora categórica *stimulus.type*. 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 que se ha planteado ha sido el siguiente:

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

En la figura XXX se muestra el *output* obtenido del modelo:

**COPIAR IMAGEN DEL OUTPUT DEL MODELO mod.co.slv2**

En la figura xxx (output del modelo), se observa que finalmente el conjunto de datos está compuesto por las variables *log(co.pre)* y *log(co.reac)*, ambas significativas y el modelo también significativo, con un valor

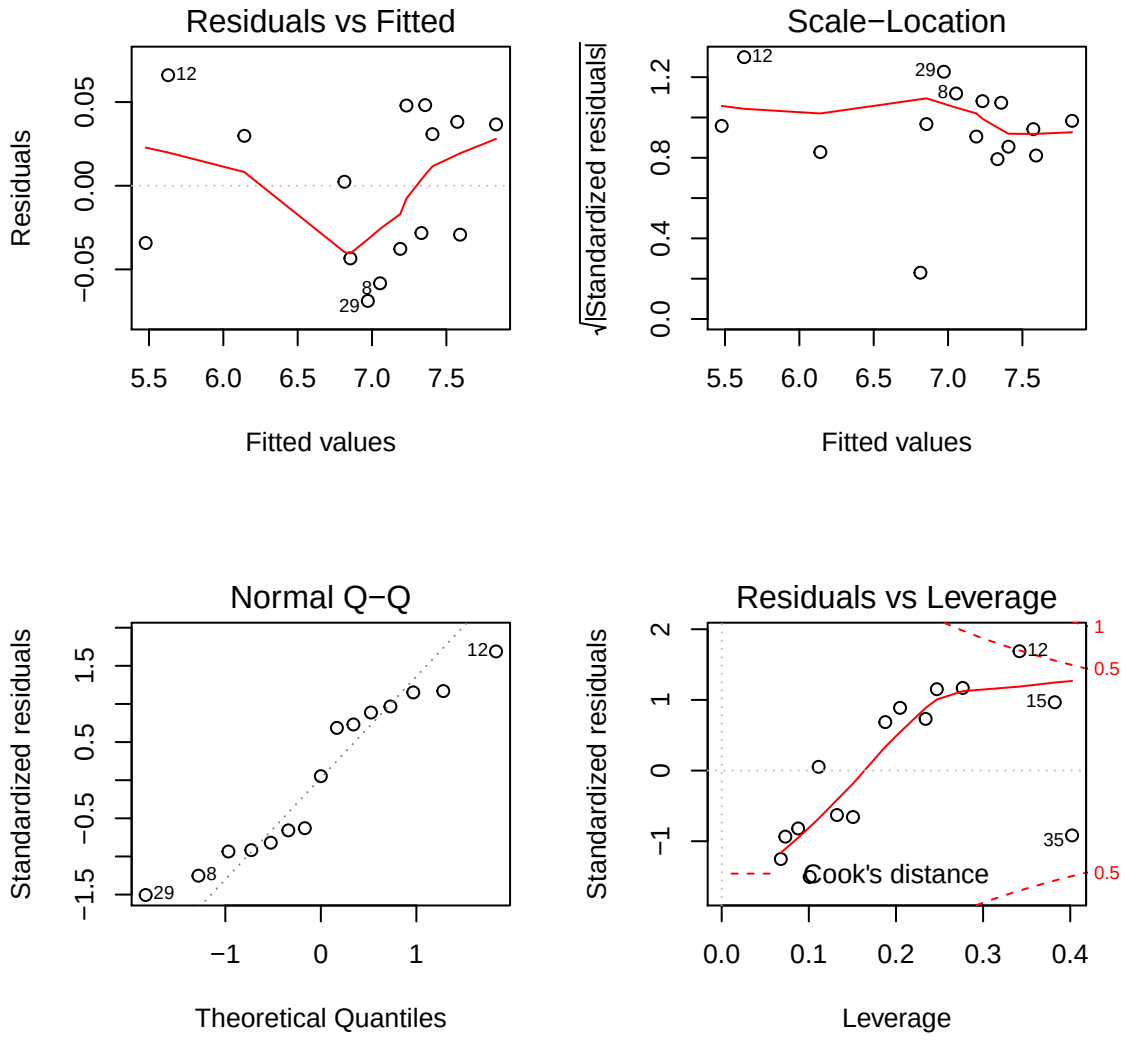
ajustado de $R^2$ muy alto.

```
Call:
lm(formula = log(co.post) ~ log(co.pre) + log(age) + stimulus.type +
    log(co.reac) + log(hr.post), data = data.co.slv)

Residuals:
      Min        1Q    Median        3Q       Max
-0.117526 -0.063807 -0.001419  0.041387  0.117501

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.21119    0.91784  -1.320    0.212
log(co.pre)     0.96757    0.04017  24.089 1.57e-11 ***
log(age)        0.04875    0.16032   0.304    0.766
stimulus.type2  0.02044    0.04595   0.445    0.664
log(co.reac)    0.25007    0.02157  11.592 7.10e-08 ***
log(hr.post)    0.18418    0.18382   1.002    0.336
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08655 on 12 degrees of freedom
  (34 observations deleted due to missingness)
Multiple R-squared:  0.9875,    Adjusted R-squared:  0.9823
F-statistic: 189.4 on 5 and 12 DF,  p-value: 5.542e-11

Start:  AIC=-83.39
log(co.post) ~ log(co.pre) + log(age) + stimulus.type + log(co.reac) +
    log(hr.post)

                Df Sum of Sq    RSS     AIC
- log(age)       1    0.0007 0.0906 -85.253
- stimulus.type  1    0.0015 0.0914 -85.096
- log(hr.post)   1    0.0075 0.0974 -83.945
<none>                       0.0899 -83.391
- log(co.reac)   1    1.0066 1.0965 -40.368
- log(co.pre)    1    4.3468 4.4367 -15.208

Step:  AIC=-85.25
log(co.post) ~ log(co.pre) + stimulus.type + log(co.reac) + log(hr.post)

                Df Sum of Sq    RSS     AIC
- stimulus.type  1    0.0013 0.0919 -87.000
- log(hr.post)   1    0.0093 0.0999 -85.493
<none>                       0.0906 -85.253
+ log(age)       1    0.0007 0.0899 -83.391
- log(co.reac)   1    1.0152 1.1058 -42.217
- log(co.pre)    1    4.4794 4.5700 -16.676

Step:  AIC=-87
log(co.post) ~ log(co.pre) + log(co.reac) + log(hr.post)

                Df Sum of Sq    RSS     AIC
- log(hr.post)   1    0.0095 0.1014 -87.227
```

```
<none>                       0.0919 -87.000
+ stimulus.type  1    0.0013 0.0906 -85.253
+ log(age)       1    0.0005 0.0914 -85.096
- log(co.reac)   1    1.0184 1.1103 -44.144
- log(co.pre)    1    4.7866 4.8785 -17.500

Step:  AIC=-87.23
log(co.post) ~ log(co.pre) + log(co.reac)

                Df Sum of Sq    RSS     AIC
<none>                       0.1014 -87.227
+ log(hr.post)   1    0.0095 0.0919 -87.000
+ log(age)       1    0.0021 0.0993 -85.599
+ stimulus.type  1    0.0015 0.0999 -85.493
- log(co.reac)   1    1.1558 1.2572 -43.907
- log(co.pre)    1    5.1664 5.2678 -18.118


Call:
lm(formula = log(co.post) ~ log(co.pre) + log(co.reac), data = data.co.slv)

Coefficients:
 (Intercept)   log(co.pre)  log(co.reac)
     -0.1716        0.9474        0.2561


Call:
lm(formula = log(co.post) ~ log(co.pre) + log(co.reac), data = data.co.slv)

Residuals:
     Min       1Q   Median       3Q      Max
-0.08718 -0.06424 -0.01347  0.04384  0.14454

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.17155    0.22866   -0.75    0.465
log(co.pre)   0.94742    0.03427   27.65 2.77e-14 ***
log(co.reac)  0.25612    0.01959   13.08 1.32e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08221 on 15 degrees of freedom
  (34 observations deleted due to missingness)
Multiple R-squared:  0.9859,    Adjusted R-squared:  0.984
F-statistic: 523.9 on 2 and 15 DF,  p-value: 1.325e-14


    Shapiro-Wilk normality test

data:  sresid2
W = 0.89059, p-value = 0.0395

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.591411, Df = 1, p = 0.44187
```

```
    studentized Breusch-Pagan test

data:  mod.co.slv2
BP = 0.8691, df = 2, p-value = 0.6476


Call:
lm(formula = log(co.post) ~ log(co.pre) + log(age) + stimulus.type +
    log(co.reac) + log(hr.post), data = data.co.slv2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.059431 -0.037334  0.007865  0.038875  0.058391

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.20947    0.60911  -0.344    0.739
log(co.pre)     0.94432    0.02756  34.270 7.57e-11 ***
log(age)       -0.03797    0.11331  -0.335    0.745
stimulus.type2 -0.02580    0.03303  -0.781    0.455
log(co.reac)    0.27564    0.02080  13.251 3.29e-07 ***
log(hr.post)    0.02421    0.12241   0.198    0.848
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05344 on 9 degrees of freedom
  (34 observations deleted due to missingness)
Multiple R-squared:  0.9963,    Adjusted R-squared:  0.9942
F-statistic: 479.3 on 5 and 9 DF,  p-value: 1.235e-10

Start:  AIC=-83.54
log(co.post) ~ log(co.pre) + log(age) + stimulus.type + log(co.reac) +
    log(hr.post)

                Df Sum of Sq    RSS     AIC
- log(hr.post)   1    0.0001 0.0258 -85.470
- log(age)       1    0.0003 0.0260 -85.350
- stimulus.type  1    0.0017 0.0275 -84.551
<none>                       0.0257 -83.535
- log(co.reac)   1    0.5016 0.5273 -40.221
- log(co.pre)    1    3.3546 3.3803 -12.351

Step:  AIC=-85.47
log(co.post) ~ log(co.pre) + log(age) + stimulus.type + log(co.reac)

                Df Sum of Sq    RSS     AIC
- log(age)       1    0.0002 0.0261 -87.330
- stimulus.type  1    0.0019 0.0277 -86.426
<none>                       0.0258 -85.470
+ log(hr.post)   1    0.0001 0.0257 -83.535
- log(co.reac)   1    0.5027 0.5285 -42.187
- log(co.pre)    1    3.4262 3.4520 -14.036

Step:  AIC=-87.33
```

```
log(co.post) ~ log(co.pre) + stimulus.type + log(co.reac)

                Df Sum of Sq    RSS     AIC
- stimulus.type  1    0.0018 0.0279 -88.307
<none>                       0.0261 -87.330
+ log(age)       1    0.0002 0.0258 -85.470
+ log(hr.post)   1    0.0000 0.0260 -85.350
- log(co.reac)   1    0.6108 0.6368 -41.389
- log(co.pre)    1    4.1601 4.1862 -13.144

Step:  AIC=-88.31
log(co.post) ~ log(co.pre) + log(co.reac)

                Df Sum of Sq    RSS     AIC
<none>                       0.0279 -88.307
+ stimulus.type  1    0.0018 0.0261 -87.330
+ log(age)       1    0.0002 0.0277 -86.426
+ log(hr.post)   1    0.0001 0.0278 -86.367
- log(co.reac)   1    0.6660 0.6939 -42.102
- log(co.pre)    1    4.2337 4.2616 -14.876


Call:
lm(formula = log(co.post) ~ log(co.pre) + log(co.reac), data = data.co.slv2)

Coefficients:
 (Intercept)   log(co.pre)  log(co.reac)
     -0.2795        0.9490        0.2767


Call:
lm(formula = log(co.post) ~ log(co.pre) + log(co.reac), data = data.co.slv2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.068843 -0.035982  0.002397  0.037403  0.066058

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -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

Residual standard error: 0.04822 on 12 degrees of freedom
  (34 observations deleted due to missingness)
Multiple R-squared:  0.9959,    Adjusted R-squared:  0.9953
F-statistic:  1472 on 2 and 12 DF,  p-value: 4.485e-15
```




En la Figura XXX, se muestra el comportamiento de los residuos del modelo, en términos de normalidad, homocedasticidad, valores *outliers* y linearidad. En términos de la linearidad, 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 influyentes (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.

- **Normalidad:**

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 XXX (Normal QQ) no parece que las observaciones sigan una distribución normal, y esto parece estar infuenciado por las pocas observaciones del conjunto de datos, la cual está compuesta por 52 observaciones tras haber eliminado los tres valores infuyentes mencionados previamente.

- **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 a los valores influyentes 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.

- **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.pre)* y *log(co.reac)* son 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.

#### 1.1.7 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 XX de este documento. Aunque la linearidad 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 (metodo Akaike) para la selección de modelos. La ecuación del modelo *mod.co.slv2* con los coeficientes es la siguiente:

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

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 Anexos

## 2.1 Anexo C - modelo cortisol 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).

### 2.1.1 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:

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

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 linearidad, 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.

### 2.1.2 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.

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

Se ha tenido que eliminar la variable *hr.post* del modelo, debido a los valores faltantes que hay en los conjuntos de datos.

### 2.1.3 Modelo III

El modelo *mod.co.p4* 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.

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

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 BoxCox 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*:

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

## 2.2 Anexo D - modelo cortisol en sangre

En el presente Anexo D 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).

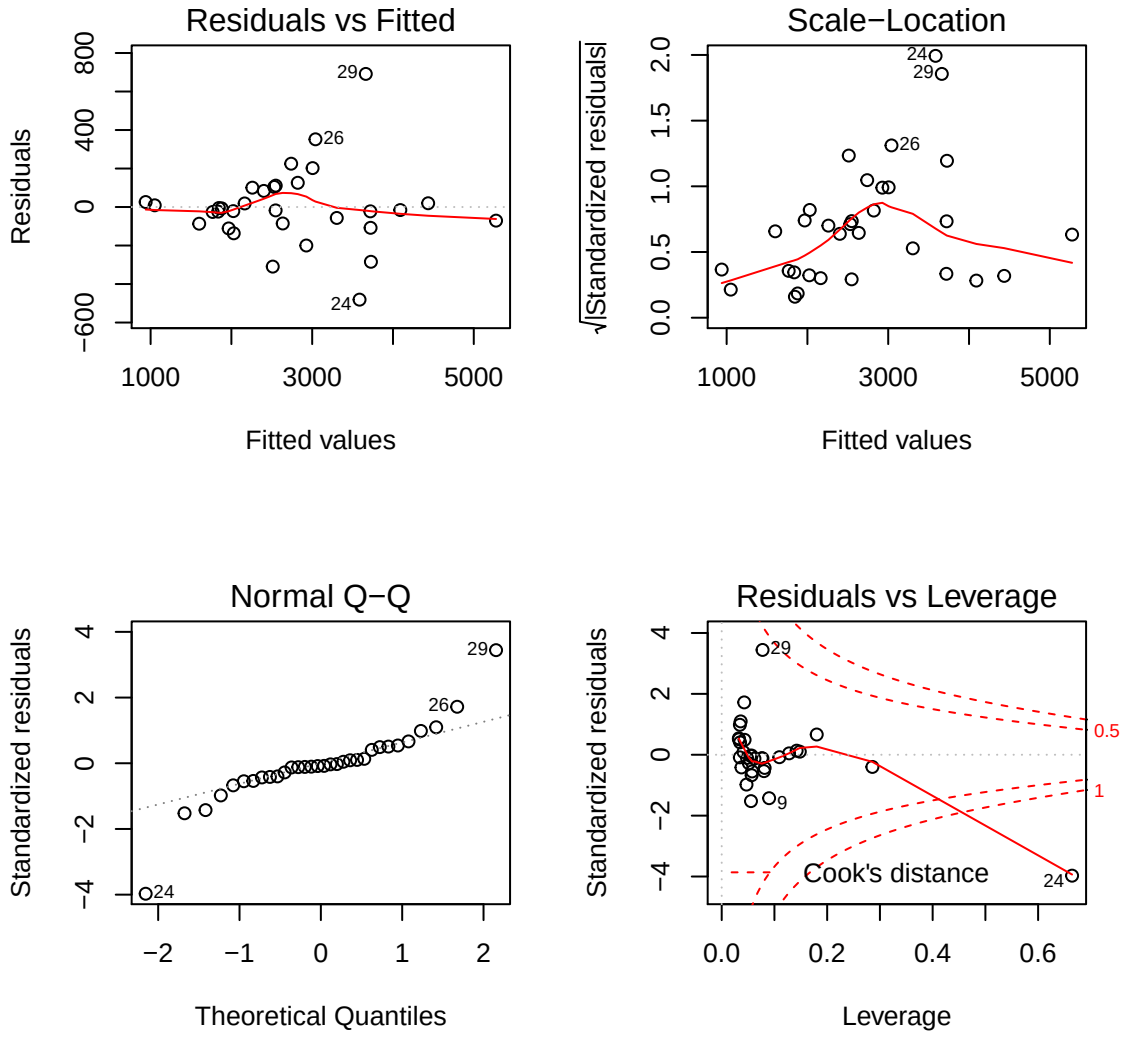
### 2.2.1 Modelo I

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:

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

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 Figrua XXX), se observa que los residuos del modelo no son homocedásticos (se ha generado una forma de *campana*) ni tampoco cumplen el supuesto de la linearidad. 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, los residuos son heterocedásticos.

**INCLUIR FIGURA RESIDUOS MODELO mod.co.sngr**




Los resultados observados en los gráficos de la figura xxx 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

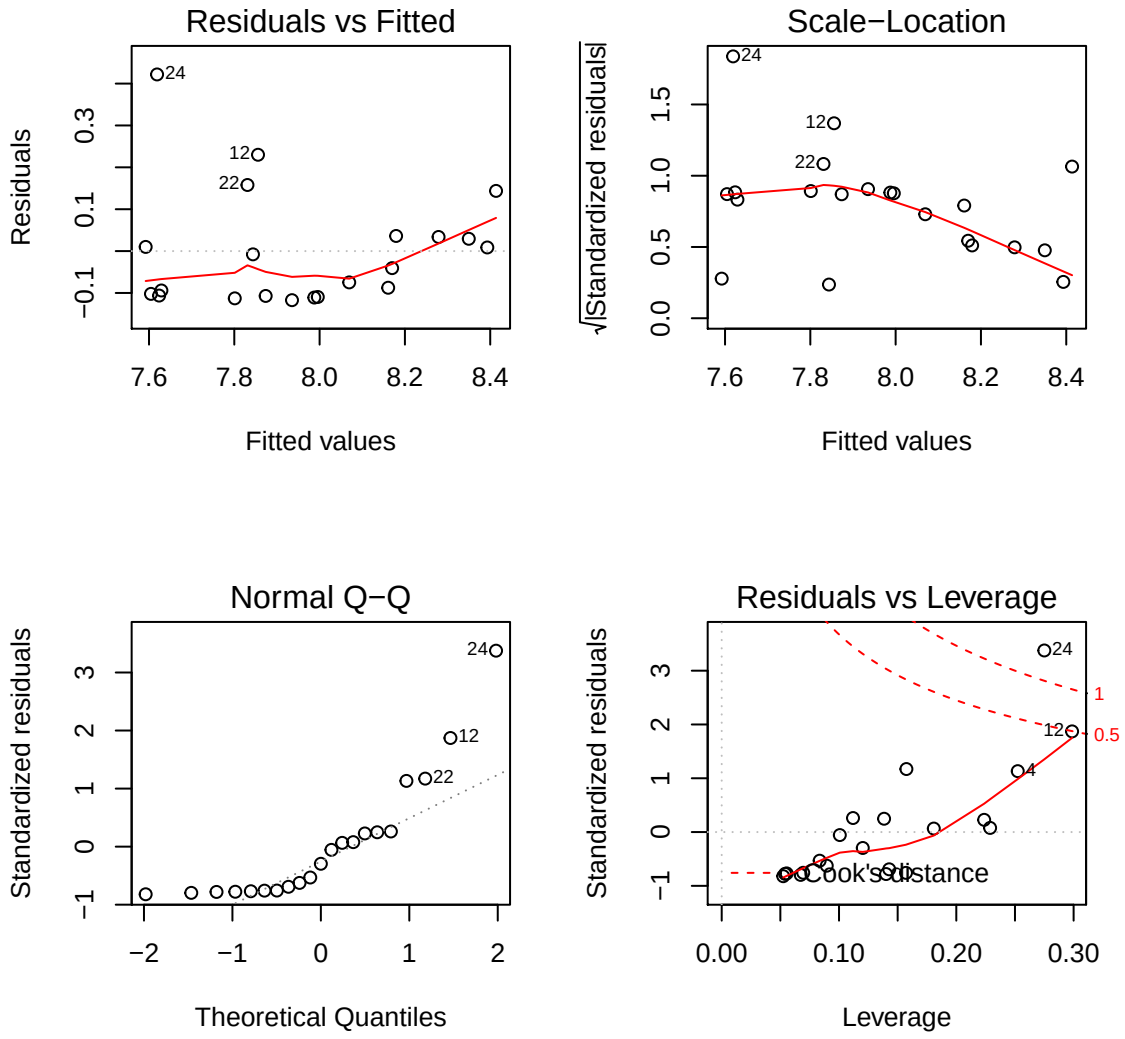
sangre.

### 2.2.2 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:

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

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 seguido con un valor del $R^2$ ajustado alto (con un valor de 0.76), pero tampoco se han cumplido los supuestos necesarios para aceptar 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:



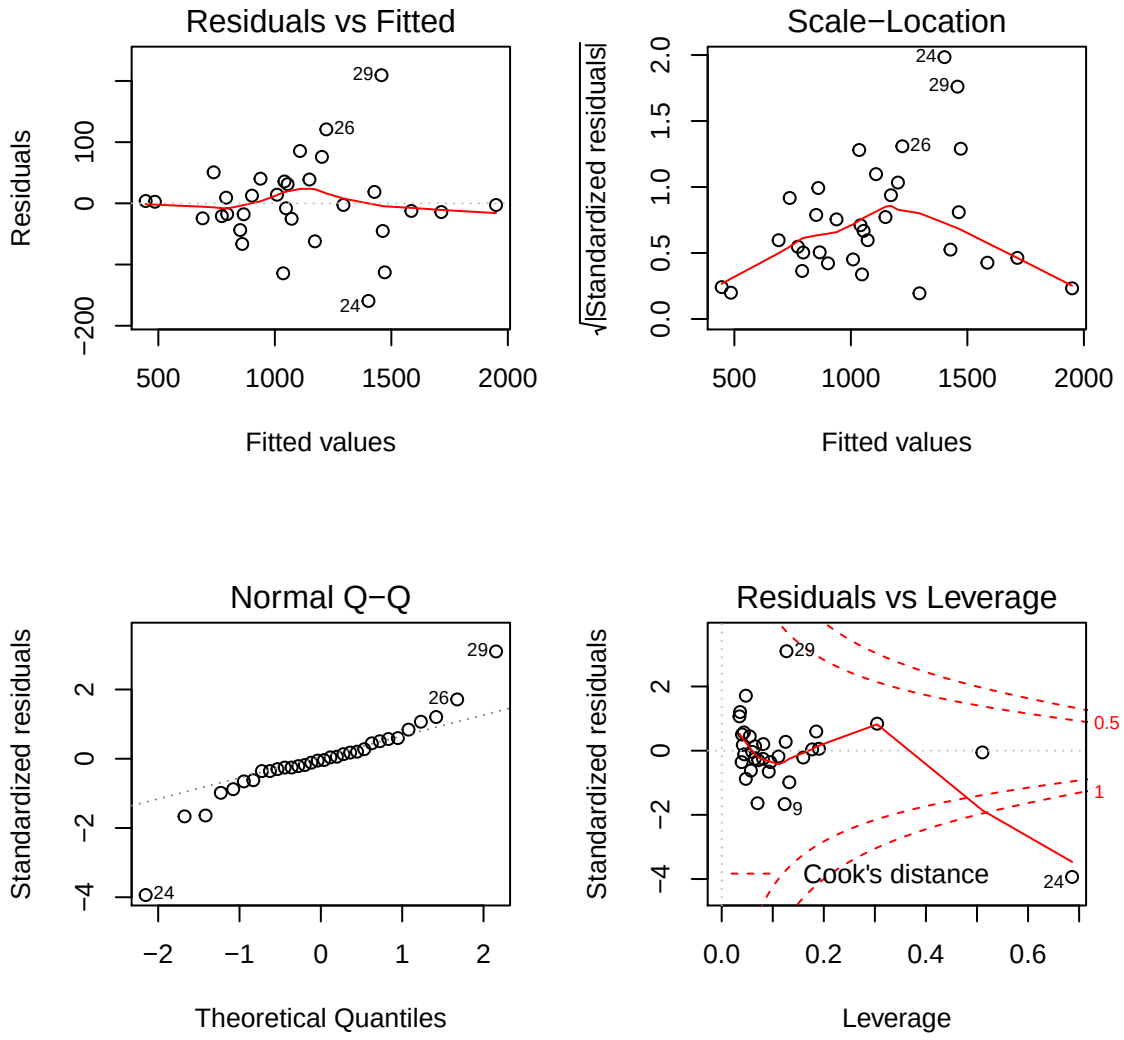
A parte de los resultados obtenido en los test, los resultados gráficos observados en la Figura XXX son suficientes para descartar el modelo *mod.co.sngr2*, ya que tampoco se cumple la linearidad de los residuos, y en el caso de la homocedasticidad, gráficamente no se aceptaría, aunque en uno de los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

### 2.2.3 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 BoxCox 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 = 0.86*, 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:

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

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 sobre la variable respuesta y la significancia en el modelo. Según este método, 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 para predecir la variable respuesta *co.post*. En la figura XXX se muestra el comportamiento de los residuos del modelo, donde se observa en el gráfico *scale location* que las varianza de los residuos no es constante debido a la forma acampanada que se genera. Sin embargo, cabe destacar que la linearidad 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.




## 2.3 Anexo E - modelo cortisol 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 BoxCox sobre la variable respuesta).

### 2.3.1 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:

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

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 eliminado 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.

### 2.3.2 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.*

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

Para ello, se transforma logarítmicamente la variable respuesta. 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, 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=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, si que se observan valores outliers, y en referencia a la linearidad del modelo, se observa que en el gráfico *Residuals vs Fitted*, se produce una parábola, lo cual muestra la falta de linearidad 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.

### 2.3.3 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 BoxCox 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 = 0.50*, 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 en un primer momento se define mediante la siguiente fórmula:

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

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. Sin embargo, el modelo no ha aceptado las hipótesis nula de normalidad, ya que el p-valor obtenido en el test de Shapiro-Wilk tiene un valor de 0.0003, ni tampoco el de la homocedasticidad, ya que ha obtenido un p-valor en ambos tests aplicados menores que 0.05. Respecto a los gráficos de los residuos, se observan bastantes observaciones *outliers*, que por ejemplo afectan a la distribución de linearidad para los valores más altos, o 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, ha obtenido un valor muy alto, descartandolo frente a los otros modelos planteados.

# References

Barrera, Monica Alejandra Mondragon. 2014. "Uso de La Correlacion de Spearman En Un Estudio de Intervencion En Fisioterapia." *Movimiento Cientifico* 8 (1): 98–104.

Martinez Ortega, Rosa Maria, Leonel C Tuya Pendas, Mercedes Martinez Ortega, Alberto Perez Abreu, and Ana Maria Canovas. 2009. "El Coeficiente de Correlacion de Los Rangos de Spearman Caracterizacion." *Revista Habanera de Ciencias Medicas* 8 (2): 0–0.

Miller, Robert, Franziska Plessow, Clemens Kirschbaum, and 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.

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.

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.