Proyecto Bioestadística

EYP3507

Nombre: Diego Aravena
Alonso Campos
Josefa Silva

Problema 1

Introducción

Se tiene información de un estudio realizado a 59 individuos que obtiene el número de convulsiones epilépticas en 4 intervalos de tiempo de duración de dos semanas. Se cuenta con información de un grupo de individuos tratados y otro de control.

La base de datos tiene información de:

- y1: número de convulsiones en el primer intervalo de 2 semanas
- y2: número de convulsiones en el segundo intervalo de 2 semanas
- y3: número de convulsiones en el tercer intervalo de 2 semanas
- y4: número de convulsiones en el cuarto intervalo de 2 semanas
- trt: indicador de tratamiento (0 control, 1 tratado)
- base: número de convulsiones en un intervalo de 8 semanas previo al estudio
- age: edad del paciente

El objetivo de este informe es estudiar la relación entre el número de convulsiones y la información sobre el tratamiento y edad de los pacientes.

Se debe tener en consideración que la base de datos no se encuentra completa, existe cierta cantidad de individuos de los cuales no se tiene registro del número de convulsiones que tuvieron en ciertos intervalos de tiempo. Debido a esto es que primero se estudiará el comportamiento de los datos faltantes y ver si considerar un método de imputación.

Análisis de los datos perdidos

La base de datos se divide entre los pacientes con datos perdidos (missing data) y los que no (complete data).

Como se presentó, las variables que se tienen a disposición corresponden al tratamiento de cada paciente, su cantidad de convulsiones antes del estudio y la edad, los cuales no poseen pérdida de datos. A continuación se presenta un resumen de las estadísticas principales de cada base de datos, aunque se debe considerar que a la variable base se le aplicó logaritmo natural con el propósito de estandarizar valores extremos.

Cuadro 1: Resumen datos

	Datos completos			Datos faltantes		
	trt	log base	age	trt	log base	age
Min	0.000	1.792	18.00	0.0	1.946	19.0
1st Qu.	0.000	2.565	24.00	0.0	2.420	21.5
Median	1.000	3.178	29.00	0.5	2.943	25.5
Mean	0.533	3.202	29.42	0.5	3.003	27.0
3rd Qu.	1.000	3.714	35.00	1.0	3.707	30.0
Max	1.000	5.017	57.00	1.0	4.025	42.0

Comparando la distribución de los datos de ambas bases se puede notar rápidamente que estos no difieren en gran medida. El tratamiento está igual de proporcionado, por lo que se descarta esta variable como variable relacionada con los datos perdidos, lo mismo con el logaritmo de la base (número de convulsiones previas al estudio), ya que estas varían más que la base original y logran mantener medias similiares. Si bien en los datos faltantes la variable edad tiende a estar ubicada para edades ligeramente menores, no se puede evidenciar un patrón significativo que permita asegurar una explicación al comportamiento de los datos perdidos.

Lo anterior se puede justificar con el hecho de que se modelaron las observaciones perdidas con respecto a las 3 covariables presentadas (trt, log base, age), de lo cual no se obtuvo ninguna relación que fuese significativa. Esto último se puede revisar en el Anexo cuya sección hace referencia a esta misma.

Lo anterior otorga una mayor flexibilidad para establecer el supuesto de que los datos se comportan de manera completamente aleatoria (MCAR).

Métodos de imputación

En esta sección se analiza el comportamiento de los datos completos luego de imputar los datos faltantes. Se seleccionaron 5 métodos distintos en base al supuesto de que los datos se asumen MCAR, veáse los cuadros 4 y 5 del Anexo , en los cuales se presentan los estimadores de la media y desviación estándar para cada uno de los métodos en conjunto con los datos previo a la imputación, es decir, los observados. Esta comparativa nace del supuesto de que los datos observados corresponden a una submuestra de los datos originales, los que supuestamente deben contener casi los mismos estimadores, de esta manera se puede estudiar cuál de todos los métodos es el que menos sesgo produce.

Las técnicas de imputación son:

- **Media**: se reemplaza por la media de la semana correspondiente de los datos observados.
- **Mediana**: se reemplaza por la mediana de la semana correspondiente de los datos observados.
- **Hot-Deck aleatorio**: se reemplaza aleatoriamente por los datos de los pacientes con características similares.
- **Hot-Deck mínima distancia**: se reemplaza por el dato cuyo paciente contenga las características más similares
- **Hot-Deck medias**: se reemplaza por la media de los datos provenientes de los pacientes con características similares.

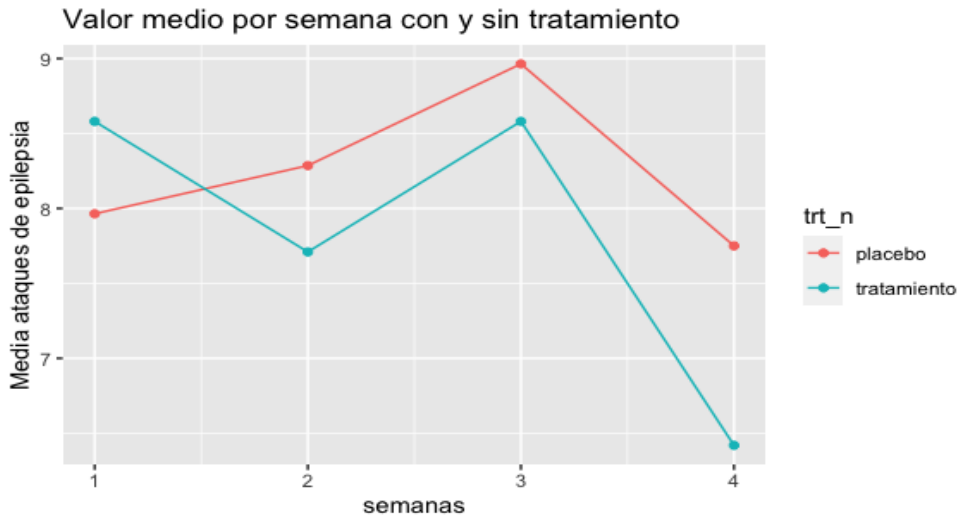
Para ser más concisos al momento de elegir qué método conviene más, a continuación se muestran los errores de estimación de cada estimador (media y desv. estándar)

Parámetro	mean	median	hotdeck al	hotdeck min	hotdeck mean
Media	0.710	0.828	0.706	0.604	0.679
Desv. estandar	1.776	1.778	1.677	1.778	1.660

Se puede observar que los errores de estimación para la media son menores en el método de imputación Hot-Deck mínima y de medias, sin embargo, viendo los errores en la desviación estándar, el que adquiere menor variabilidad en la estimación es el método Hot-Deck de medias. Por ende, este último es el método que se utilizará para el posterior análisis entre la relación del número de convulsiones y el tratamiento/edad en cada paciente.

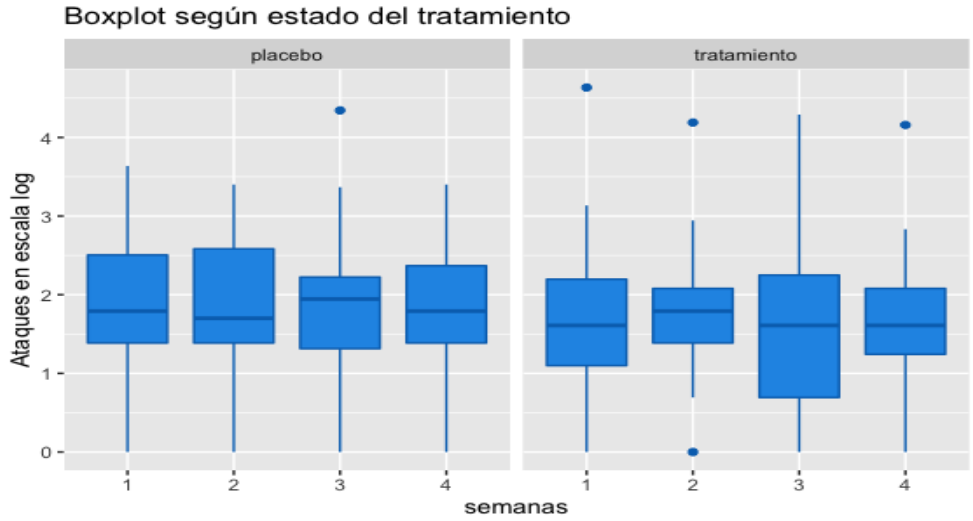
Análisis exploratorio

Como primer acercamiento a los datos, a continuación se presentan los promedios del número de convulsiones por semana durante todo el estudio distinguiendo por tratamiento.



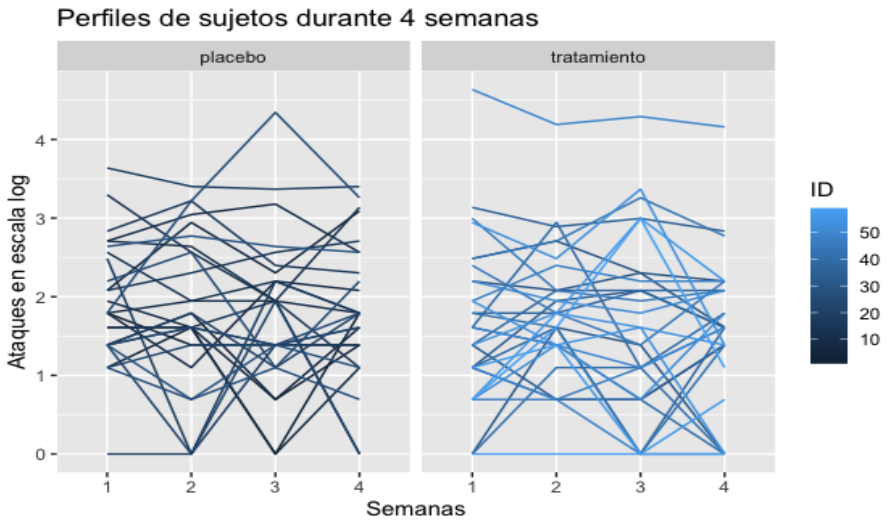
Rápidamente se puede observar que el grupo tratado empieza peor que el grupo placebo, no obstante, a lo largo del estudio este grupo va disminuyendo su promedio de ataques. Por otro lado, el grupo placebo no presenta cambios significativos después de las 4 semanas.

Debido a la gran diferencia de escala que existe en las variables edad, y1, y2, y3, y4 y base, es que se decidió transformarlas mediante escala logarítmica para aportar estabilidad en los regresores y reducir las observaciones atípicas. A continuación se muestra el comportamiento de cada grupo por semana.



Se observa una muy leve variación entre los grupos que obtuvieron tratamiento y los que no en cuanto a la variabilidad de los ataques.

Posterior a ello se analizará el comportamiento de los sujetos a lo largo del estudio separando por tratamiento y manteniendo la escala logarítmica para la respuesta.



En principio no se logra percibir un patrón claro que permita afirmar de que existe una diferencia en el número de convulsiones al comparar los tratamientos. Sin embargo, lo que queda claro es que en los tratados existe un mayor número de sujetos que no tienen (o tienen muy poco) ataques a la cuarta semana, además se puede observar que los que no reciben tratamiento son más jóvenes. En el siguiente cuadro se muestra el promedio y la variabilidad del número de convulsiones que sufrieron los sujetos en cada semana, también distinguiendo a los tratados y los no tratados.

Cuadro 2: Resumen datos

Tratamiento	semana	Cantidad	Media	Varianza	Desviación Estándar
placebo	1	28	7.964	64.70	8.044
placebo	2	28	8.286	66.66	8.164
placebo	3	28	8.964	213.81	14.622
placebo	4	28	7.750	60.19	7.758
tratamiento	1	31	8.581	332.72	18.241
tratamiento	2	31	7.710	134.81	11.611
tratamiento	3	31	8.581	196.58	14.021
tratamiento	4	31	6.419	125.72	11.212

Como primera observación se puede identificar que el grupo placebo logra mantener una media semi-estable, aunque su nivel de dispersión en los datos es muy alto, esto se debe principalmente a la poca cantidad de observaciones que se tiene y a los demás factores que pueden estar influyendo. Por otro lado, sucede algo similar con la media de los sujetos que fueron tratados, si bien esta tiene una leve disminución en la cuarta semana (tal como se vió anteriormente), la variabilidad en los datos es aún más grande. Esto se puede deber a los diferentes efectos secundarios que genera el tratamiento en cada sujeto, lo que termina beneficiando a algunos y a otros no tanto, lo que trae consigo nuevas inquietudes tales como el poder identificar cuáles son los factores que determinan si el tratamiento es efectivo o no.

Tal como se menciona, algunas de las varianzas son considerablemente mayores que las medias correspondientes, por lo que para una variable de Poisson la dispersión excesiva puede ser un problema, por lo que se tendrá que probar distintos modelos.

Modelamiento

Al ajustar un modelo lineal generalizado (GLM) considerando sólo la variable *tratamiento* como variable explicativa, resultó que esta no era significativa para explicar la cantidad de convulsiones (Modelo 1). Por otro lado, al considerar las variables *edad* y *base* por separado, estas si resultaron ser significativas. (Modelo 2 y Modelo 3)

También se consideró el caso para efectos aleatorios a nivel de sujeto, pero el ajuste del modelo empeoraba aún cuando considerábamos las variables en escala logarítmica. (Modelo 4)

Posterior a ello se modificó el modelo anterior (GLM), considerando la variable *base* que es la cantidad de convulsiones que tiene cada sujeto en un intervalo de 8 semanas antes de iniciar el tratamiento, la *edad* y una variable nueva llamada *semana*, que es un indicador de cada periodo del estudio, en este caso consta de 4 periodos de 2 semanas cada uno.

De los distintos modelos ajustados se llegó a la conclusión que al ajustar por variable *base*, el *tratamiento* es significativo y de los 4 periodos, el último era el único significativo. Como las respuestas fueron conteos durante un periodo de dos semanas, queremos interpretar los resultados en términos de efecto en una tasa semanal, por lo que se tienen dos soluciones:

1. Dividir los valores por dos para obtener una tasa semanal, sin embargo, este valor puede ser decimal y la distribución Poisson sólo recibe valores enteros.

2. Incluir en nuestro modelo la variable *offset(logtime)*, la cual se utiliza para indicar el periodo de la exposición en la regresión Poisson, en este caso sería *offset(log(2))*, pues cada periodo consta de dos semanas. (Modelo 5)

Debido que los datos poseen varianzas considerablemente mayores que las medias correspondientes, se decidió explorar distintos modelos y verificar si los resultados obtenidos con el método GLM eran consistentes, ya que, para una variable Poisson tener una dispersión excesiva puede ser un problema, pues, la media y la varianza deben ser iguales o parecidas.

A continuación se muestra un resumen de los estimadores obtenidos en cada modelo utilizado:

	GLM	GEE-Indep(1)	GEE-Exchg(1)	GEE-Exchg(free)
(Intercept)	-0.036	-0.036	-0.035	-0.035
	(0.118)	(0.293)	(0.296)	(0.296)
trt_ntratamiento	-0.161 ***	-0.161	-0.157	-0.157
	(0.047)	(0.158)	(0.158)	(0.158)
semana4	-0.168 **	-0.168 *	-0.168 *	-0.168 *
	(0.056)	(0.067)	(0.067)	(0.067)
age	0.019 ***	0.019 *	0.019 *	0.019 *
	(0.003)	(0.009)	(0.009)	(0.009)
base	0.023 ***	0.023 ***	0.023 ***	0.023 ***
	(0.001)	(0.001)	(0.001)	(0.001)

Y como era de esperar, debido a la sobredispersion de los datos el tratamiento para los modelos creados con el método GEE (Generalized Estimation Equation) no es significativo y baja la significancia de la edad.

Ahora, ¿Cómo se puede arreglar esto?. Agregando la interacción que puede tener la base y el tratamiento en conjunto se concluye que estos poseen un alto impacto, además cambiando las variables edad y base en su escala logarítmica, a simple vista se nota una mejoría en todos los ajustes:

	GLM	GEE-Indep(1)	GEE-Exchg(1)	GEE-Exchg(free)
(Intercept)	-4.496 ***	-4.496 ***	-4.490 ***	-4.490 ***
	(0.394)	(0.791)	(0.794)	(0.794)
logbase	0.950 ***	0.950 ***	0.950 ***	0.950 ***
	(0.044)	(0.095)	(0.097)	(0.097)
trt_ntratamiento	-2.095 ***	-2.095 **	-2.095 **	-2.095 **
	(0.247)	(0.687)	(0.689)	(0.689)
logage	0.804 ***	0.804 ***	0.803 ***	0.803 ***
	(0.106)	(0.211)	(0.211)	(0.211)
semana4	-0.168 **	-0.168 *	-0.168 *	-0.168 *
	(0.056)	(0.067)	(0.067)	(0.067)
logbase:trt_ntratamiento	0.554 ***	0.554 **	0.554 **	0.554 **
	(0.064)	(0.182)	(0.182)	(0.182)

Para el enfoque GEE, se utilizan dos tipos de estructura de correlación, independiente (Las mediciones repetidas no están correlacionadas) o intercambiable (todos los pares de observaciones en la misma unidad tienen una correlación común) en donde se restringe el parámetro de escala para que sea uno (GEE-INDEP(1) y GEE-EXCHG(1)), sin embargo, para estos datos existe una sobredispersión (la varianza es mayor que el valor medio) que debe adaptarse, permitiendo que este parámetro se estime libremente (GEE-EXCHG(FREE)).

Al presentar coeficientes tan significativos y con valores p casi perfectos (Modelo 6) se decidió visualizar el problema de la sobredispersión de forma cuantitativa, calculando el parámetro de dispersión, el cual resulto ser 4.67. Lo anterior alerta sobre la importancia del problema de sobredispersión dentro del modelo. Se comprueba cuánto se ven afectadas las estimaciones de los coeficientes por la sobredispersión (Modelo 7), en el caso de el modelo en el cual no se considera la interacción entre base y tratamiento y no se usa la escala logaritmica, los coeficientes significativos disminuyen a la mitad. El caso en el cual se considera la interacción entre base y tratamiento usando la escala logaritmica, la última semana del estudio se vuelve insignificante y baja el valor p de los coeficientes, lo que cambia toda la interpretación del modelo.

Por lo tanto, se decidió abordar este problema de dos formas:

Permitir estimación de dispersión

Una forma de solucionar el problema de la sobredispersion es estimar el parámetro de dispersión dentro del modelo. Esto se puede hacer a través de las cuasi-familias, en este caso, sería una cuasi-poisson. Este procedimiento nos dice que tres de los coeficientes de los predictores son significativos (Modelo 8).

Reemplazar la distribución Poisson con Binomial Negativo

Otra forma de abordar la sobredispersión en el modelo es cambiar el supuesto de distribución al binomial negativo en el que la varianza es mayor que la media . Como la relación entre la desviación sobre los grados de libertad es grande, es decir, la razón de una estimación a su error estándar es mayor a uno (2.73), este modelo no se ajusta bien. (Modelo 9)

Conclusión

A lo largo de esta sección se probaron diversos modelos con distintos métodos, si bien el tratamiento resulta significativo en algunos de los modelos realizados anteriormente, hay que recalcar que solo sucede si se considera la variable base, ya sea en escala logarítmica o no. Sin embargo la variable edad resulta ser significativa por sí sola, es decir, posee un impacto consistente sobre la cantidad de ataques epilépticos que tiene cada sujeto. También se consideró la sobredispersión de los datos, lo que da indicio de que al utilizar una familia poisson para generar los modelos no viene siendo lo más óptimo para este estudio, por lo tanto al modificar la distribución en el modelo final (Modelo 10) por la familia cuasi-poisson se estaría teniendo en consideración este factor al momento de sacar conclusiones.

Problema 2

Introducción

Se tiene información de un grupo de 5000 personas expuestas a cierto virus. Dicha información se desgloza en las siguientes variables

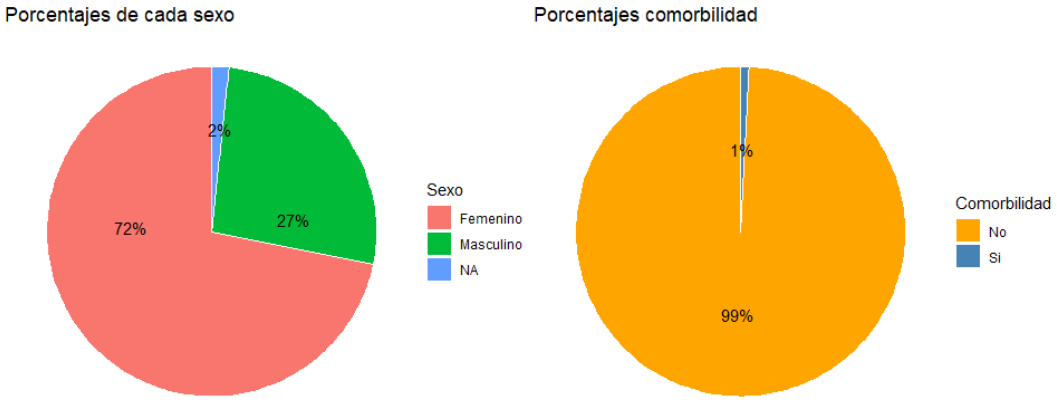
- **fecha**: fecha de hospitalización
- **edad**: edad del paciente
- **sexo**: sexo del paciente
- **comorbilidades**: si el paciente tiene comorbilidades (1) o no (0)

La información fue recabada entre un período del 04 de Abril del 2020 y el 01 de Agosto del 2021.

A continuación se estudiarán los factores de riesgo de caer hospitalizado en base a la información obtenida.

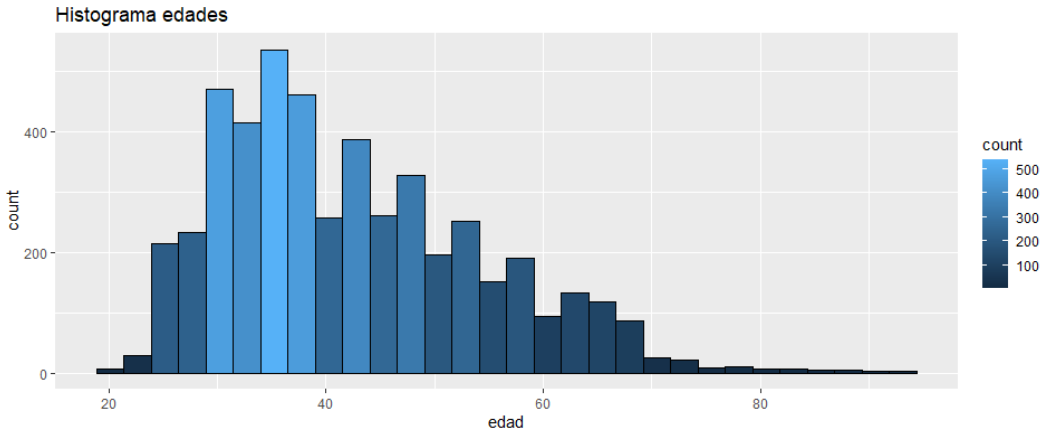
Análisis exploratorio

Lo primero a analizar es el comportamiento de los datos, para ello se parte revisando los porcentajes que hay para cada sexo y para comorbilidad. A continuación se muestran los gráficos de torta para cada caso.

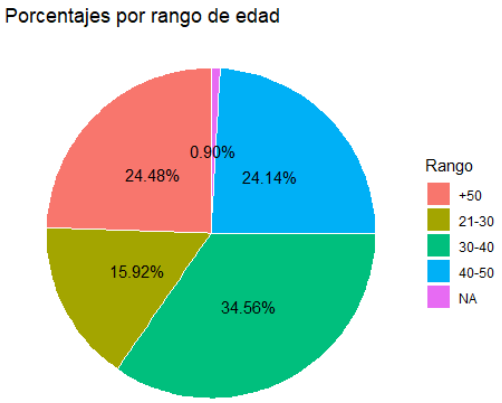


Claramente se puede notar una mayor proporción de mujeres que de hombres, además se debe considerar que de los 5000 sujetos un 2 % no presenta su sexo. Por otra parte, tan solo un 1 % presenta comorbilidades, lo que quiere decir que el 99 % de los sujetos tiene un estado normal de salud.

Luego se analiza el comportamiento de las edades de cada sujeto, esta información se presenta en el siguiente histograma

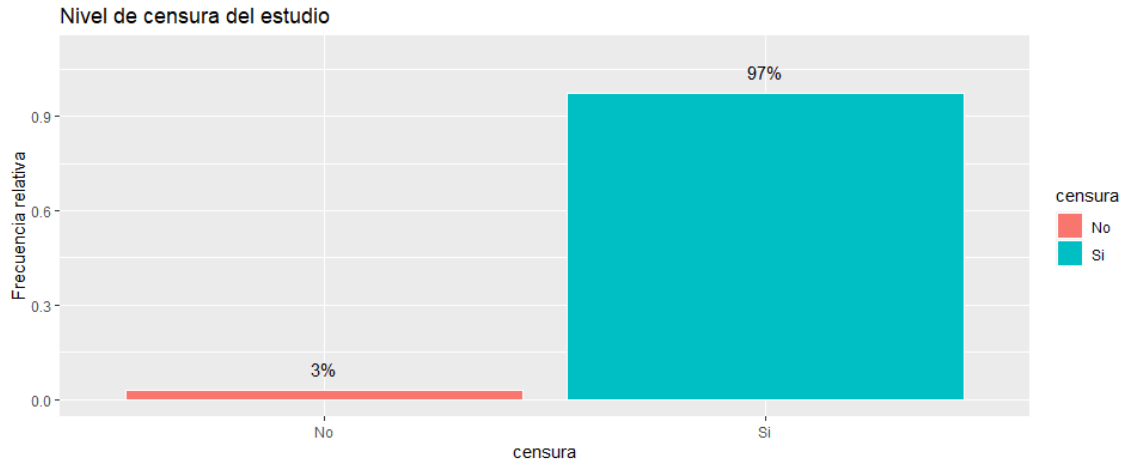


Se puede ver que hay de todas las edades, lo cual es beneficioso para comprender mejor los efectos que pueda tener el virus y así identificar a las personas con mayor riesgo en base a esta variable. También se puede notar una asimetría positiva en la distribución de los datos, recolectando una menor cantidad para edades más avanzadas, además entre las edades 30 y 40 se puede notar un aumento en la cantidad de datos que en cualquier otro rango. Por esto, ahora interesa conocer cómo se distribuyen las edades entre sí, para esto se presenta el siguiente gráfico de torta:



Tal como se había sospechado, la gran mayoría de los sujetos tiene entre 30 y 40 años de edad, los cuales representan un 34.87 % del total. Se puede ver además, que la cantidad de sujetos con edades superiores a 50 logran asimilar la cantidad de sujetos con edades de entre 40 y 50 años, lo que deja a los sujetos menores de 30 años como los con menor frecuencia.

Por último a continuación se muestran los niveles de censura ocurridos durante el estudio



Tan solo se llega a contabilizar un 3 % del total de sujetos que caen en hospitalización, lo que quiere decir que el nivel de censura es tan alto que un 97 % de los sujetos no cayó hospitalizado producto del virus durante el período de estudio.

Técnicas de imputación

- **Eliminación de los datos**: En primera instancia se eliminan todos los datos que poseen al menos un valor nulo en las variables distintas a `fecha`, es decir, las variables `edad`, `sexo` y `comorbilidades`. Al utilizar esta técnica, se eliminan 84 registros, que representan un 1,68 % del total de datos, por lo que la nueva base de datos la componen 4.916 observaciones.

- **Imputación por la media de los datos completos:** El segundo procedimiento fue imputar datos por la media (o moda, respectivamente) por los datos que poseen todas sus observaciones. Para la variable `edad` se imputa la media, en el caso de `sexo` la moda. En los casos de `comorbilidades` no se imputan valores, pues en esta variable no tiene registros nulos.

- **Imputación por la media:** En una tercera instancia se imputaron valores tomando en cuenta que poseen alguna información, es decir, anteriormente se eliminaba si el registro tenía al menos un valor faltante. Ahora, dicho registro se imputará el dato faltante, pero también aportará información con las demás variables.

- **Hot-deck**: Finalmente, se utilizó el método Hot-deck. En este caso, se eliminaron las variables que tenían como dato faltante la edad, puesto que las personas que no registraban esa variable además no reportaban su sexo. Así, al intentar imputar por este método, sería demasiada la variabilidad, pues también se sabe que ese grupo de personas no tienen comorbilidad. Entonces, al tener pocas variables para comparar, se decidió eliminar esos registros. Es importante mencionar que son un total de 45 registros que representan un 0,9 % de la base de datos original.

Luego de implementar cada uno de los procedimientos, en la siguiente tabla se resumen los estadísticos de interés de cada variable

Cuadro 3: Resumen datos

	Edad		Sexo		Comorbilidad	
	Media	Desviación Estándar	Proporción Femenino	Proporción Masculino	Proporción con comorbilidad	Proporción sin comorbilidad
Eliminación	42.4795	12.1660	0.7303	0.2697	0.0018	0.9982
Imputación media	42.4884	12.1653	0.7348	0.2652	0.0022	0.9978
Imputación Media 2	42.4884	12.1653	0.7348	0.2652	0.0022	0.9978
Hot Deck	42.4928	12.2204	0.7312	0.2688	0.0022	0.9978

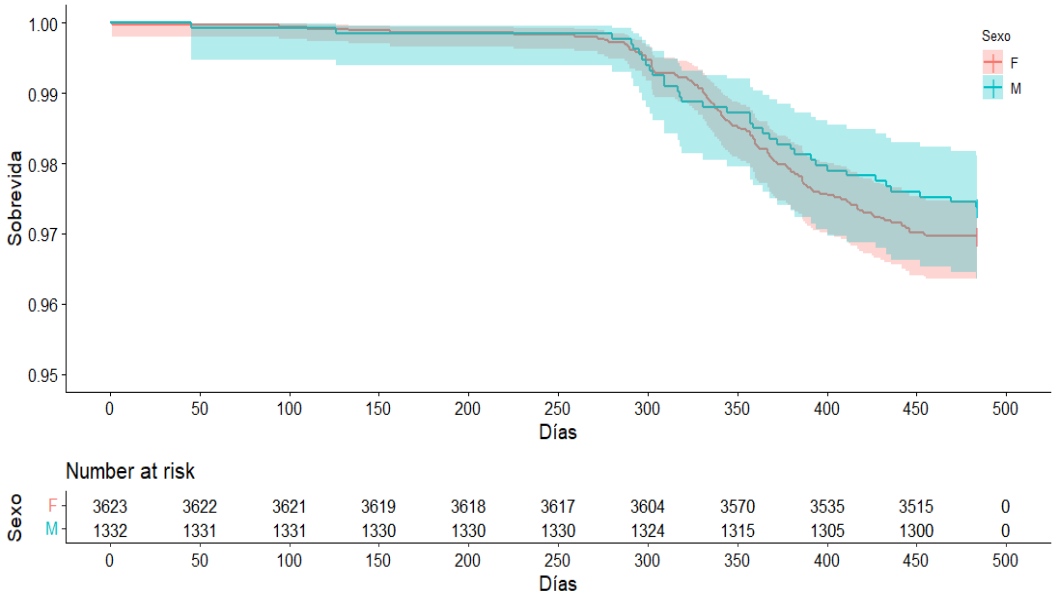
De la tabla anterior se puede observar que los valores de cada uno de los estadísticos no cambian significativamente. Esto se debe, que la proporción de daros faltantes es muy poca en proporción al total de datos. Por esta razón, cualquier método que se elija no debería impactar en los resultados. Para el análisis posterior se utilizará el método Hot Deck pues tiene una mayor estimación en la desviación estándar de la variable edad.

Kaplan Meier

En esta sección se analizarán las estimaciones de Kaplan Meier para las curvas de sobrevivencia para cada una de las variables. En particular, se abordarán dos puntos, el gráfico de las curvas estimadas de sobrevivencia de Kaplan Meier y el test Log-Rank para determinar si dichas curvas son iguales o distintas.

Variable sexo:

Luego de reemplazar los datos faltantes para la variable `sexo` por el método Hot Deck, se estiman las curvas de sobrevivencia para cada género.

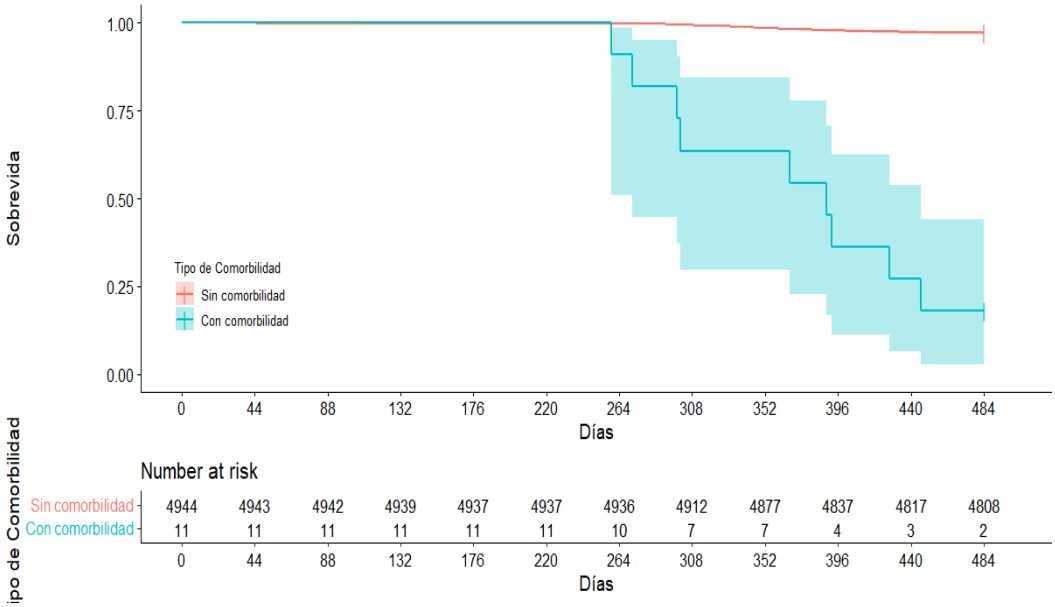


De la figura anterior es importante destacar que el eje y varía de 0,95 a 1. También, es posible apreciar que no hay diferencias significativas entre las dos curvas, considerando la magnitud del eje. Entonces, se concluye que, gráficamente, no hay diferencias significativas en cuanto al sexo del paciente.

Además, también se realiza el test log-rank para determinar, desde una perspectiva objetiva, si las dos curvas de sobrevivencia son iguales. Donde se obtiene un valor-p=0,5 (ver anexo) por ende no se rechaza la hipótesis nula. Recordemos que la hipótesis nula del test es que ambas curvas de sobrevivencia son iguales, por ende no se rechaza.

Variable comorbilidades:

De manera similar, se puede estudiar las curvas estimadas de la función de sobrevivencia en el siguiente gráfico

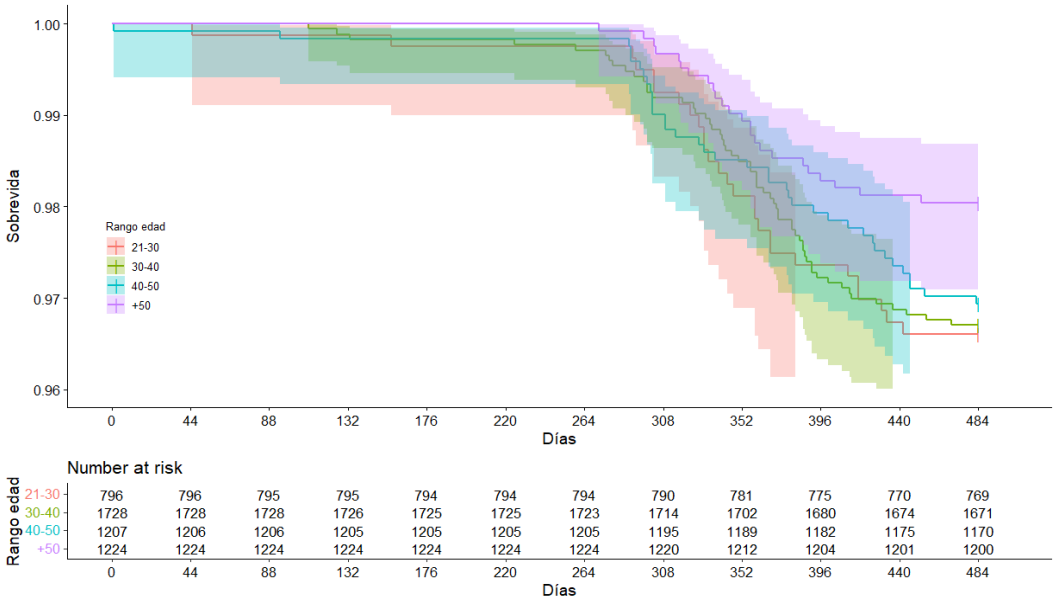


Se puede observar que, si estudiamos la sobrevivencia, considerando el tiempo hasta ser hospitalizado, por la variable comorbilidad, sí existen diferencias en las curvas de sobrevivencia. Pues se puede apreciar que la curva roja casi no tiene variaciones y se mantiene cerca de 1, mientras que las personas que sí presentan esta condición su curva varía mucho más llegando hasta el final del estudio a un valor de 0,25.

Además, también se realiza el test Log-rank, obteniendo un valor-p muy cercano a 0 (ver anexo), por lo que se rechaza la hipotesis nula y se concluye que las curvas son diferentes.

Variable edad:

Para poder comparar las funciones de sobrevivencia, se crearon 4 categorías, que representan rangos de edad, éstas son desde 21 hasta 30 años, de 31 a 40 años, entre 40 y 50 años y mayores a 50 años. Después se realiza el gráfico de cada una de las curvas.



De la figura anterior se puede observar que en algunas categorías presentan leves diferencias, por ejemplo, la primera y segunda categoría. Sin embargo, otras como la primera y la última se aprecian diferencias más notorias.

Utilizando el test Log-Rank obtenemos un valor-p igual a 0,1 (ver anexo), por lo que no se rechaza la hipótesis nula. En este caso, como son más de dos categorías la hipótesis nula será que no hay diferencia entre las curvas de los grupos. Por ende, se concluye que no existen diferencias significativas entre las curvas de los datos agrupados por rango de edad.

Modelo de Cox

En esta sección se estudiarán los factores de riesgo utilizando el modelo de Cox.

En primer lugar, se ajusta el modelo en `R`, donde se obtienen los siguientes resultados

	Estimación	Estadístico z	Valor-p	Exp	Limite Superior	Limite inferior
Edad	-0,0237	-3,134	0,0017	0,9766	0,9623	0,9912
Sexo (Masculino)	-0,0918	-0,471	0,6377	0,9123	0,6227	1,3366
Comorbilidad	4,2037	11,818	$\leq 2 \cdot 10^{-16}$	66,9350	33,3331	134,4098

Se puede apreciar que el valor-p de la variable `sexo` (utilizando a masculino como variable dummie) es mayor a 0,05 por lo que dicha variable no es significativa, esto se puede comprobar pues al utilizar la exponencial del predictor asociado a `sexo` y el intervalo de confianza, este último contiene el valor 1.

Por otro lado, las demás variables, tanto edad como comorbilidad sí son significativos, es decir, el valor-p asociado es menor a 0,05 y utilizando la exponencial del predictor, los intervalos de confianza no contienen el valor 1.

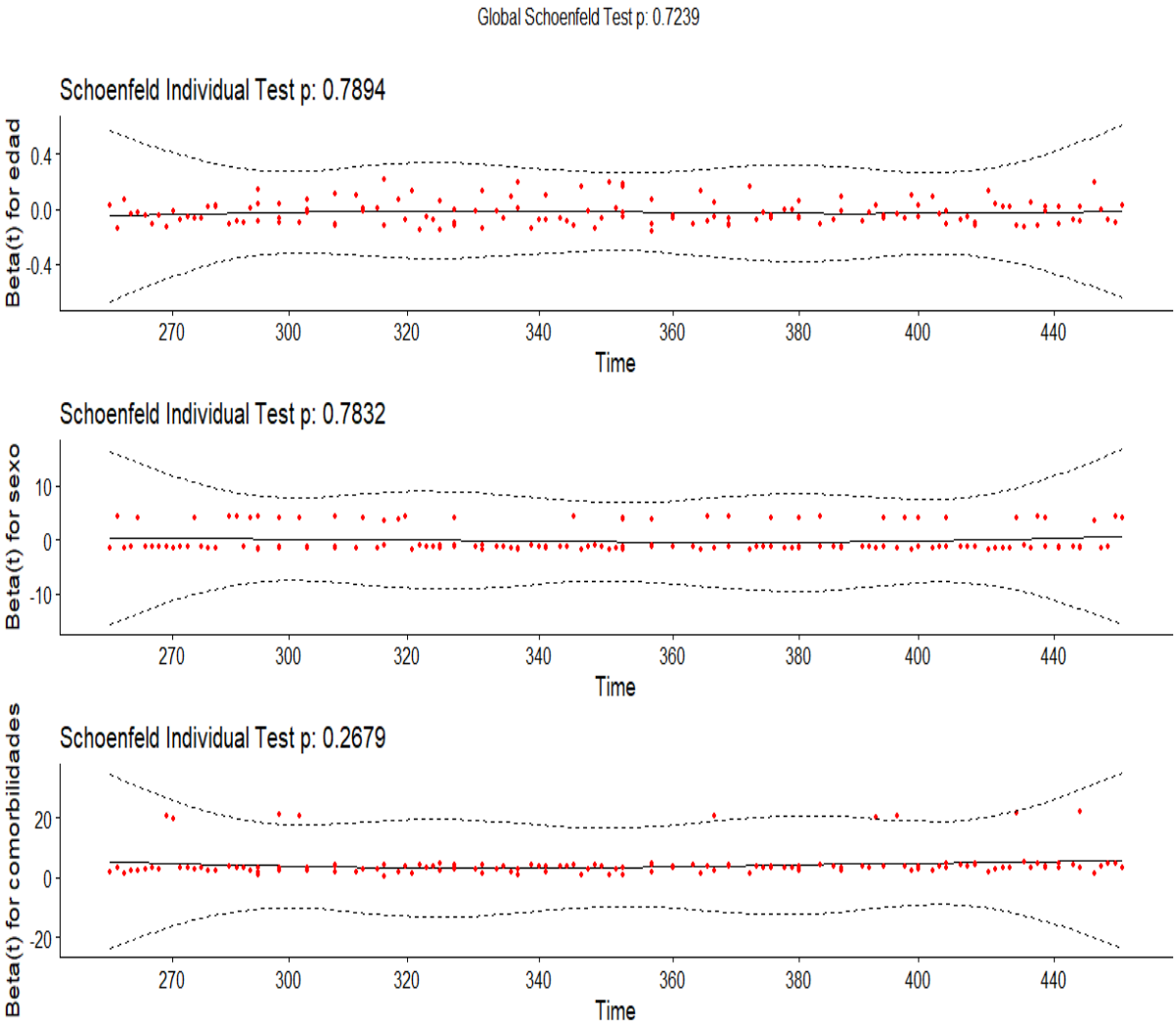
Analizando los valores del predictores, específicamente, el valor de e^{β} podemos señalar que, la tasa de hospitalización de un paciente es 0,97 veces mayor que un paciente con un año más, también se puede expresar como la tasa de hospitalización de un paciente es 1,023 veces mayor para un paciente con un año menos. De la misma manera, la tasa de hospitalización de las mujeres es 1,09 veces mayor que el de los hombres, cabe destacar que este valor no es significativo. Finalmente, en la variable c`comorbilidades`, la tasa de hospitalización de las personas con comorbilidades es 66,93 veces que el de las personas sin esta condición.

Posteriormente, se evalúa si el modelo de riesgos proporcionales es el adecuado. Para esto utilizamos la función `cox.zph`, donde se obtienen los siguientes resultados

	Estadístico	Grados Libertad	Valor-p
Edad	-0,0714	1	0,79
Sexo	0,0757	1	0,78
Comorbilidad	1,2274	1	0,27

Con lo anterior, se puede apreciar que ninguna de las variables en el test tiene un valor-p menor que 0,05, como la hipótesis nula asume el supuesto, entonces se tiene evidencia para no rechazarla, por lo que es razonable asumir este supuesto en el modelo.

En este sentido, para analizar el supuesto de riesgos proporcionales, analizaremos los residuos de Schoenfeld



En la figura anterior se puede evidenciar que en cada uno de los gráficos de las variables, no hay patrones en el tiempo. De esta manera, evidenciamos que las variables no dependen del tiempo.

Conclusión

En esta sección, en primer lugar, se estudió la naturaleza de los datos de cada una de las variable. Luego se analizaron los datos faltantes, para cada una de las variables y diferentes métodos de imputación.

Posteriormente, se analizó las curvas de sobrevivencia para cada variable, en el caso de edad se categorizó. Donde se concluye que en edad y comorbilidades hay diferencias significativas en la estimación de dichas curvas.

También, se utilizó el modelo de riesgos proporcionales de Cox donde los principales hallazgos fueron que el género no es significativo, además que si bien la edad es significativa, para cada año de diferencia entre los pacientes la tasa de hospitalización no es muy alta, en comparación con la comorbilidad. En ésta última variable, se concluye que es significativa y además la tasa de hospitalización entre las personas que presentan comorbilidad es muy alta en comparación a las que no la presentan.

En síntesis, el principal factor para caer hospitalizado es la comorbilidad, así se recomienda estudiar o invertir más dinero, pues representan una minoría (alrededor de un 1 %) y tienen un riesgo mucho mayor que las personas que no tienen comorbilidad. En segundo lugar, habría que estudiar la población joven pues con los datos disponibles se concluye que éstos son más propensos a estar hospitalizados. Finalmente, no se puede concluir un mayor riesgo según el sexo del paciente.

Anexos

Problema 1

Análisis de datos faltantes

Mediante el método stepwise se busca obtener el mejor modelo que explique la variable obs (1 si se tiene dato perdido, 0 sino) teniendo en cuenta solo las variables trt, age y base

```
base_model <- glm(obs ~ 1, data = epilepsia,
                  family = binomial(link = "cloglog"))
full_model <- glm(obs ~ trt+age+base, data = epilepsia,
                  family = binomial(link = 'cloglog'))
step(base_model,
     scope=list(lower=formula(base_model),upper=formula(full_model)),
     direction="both", trace = T)

#  Start:  AIC=66.66
# obs ~ 1
#
#        Df Deviance    AIC
# <none>      64.656 66.656
# + age   1   63.472 67.472
# + base  1   63.738 67.738
# + trt   1   64.609 68.609
#
# Call:  glm(formula = obs ~ 1, family = binomial(link = "cloglog"),
#            data = epilepsia)
#
# Coefficients:
# (Intercept)
#      -1.306
#
# Degrees of Freedom: 58 Total (i.e. Null);  58 Residual
# Null Deviance:            64.66
# Residual Deviance: 64.66      AIC: 66.66
```

En base a este resultado y al posterior chequeo de cada regresión simple para la variable obs, se concluye que ninguna de las variables logra explicar significativamente el comportamiento de los datos perdidos.

Métodos de imputación

Cuadro 4: Comparación de medias para datos imputados

		Media datos imputados				
	Media observada	mean	median	hotdeck al	hotdeck min	hotdeck mean
Semana 1	9.266667	8.271186	8.135593	8.322034	8.677966	8.322034
Semana 2	8.622222	8.474576	8.135593	8.237288	8.118644	8.169491
Semana 3	7.022222	8.796610	8.457627	8.559322	8.491525	8.576271
Semana 4	7.622222	7.271186	7.169492	7.186441	7.203390	7.322034
Total	32.533333	32.813559	31.898305	32.305085	32.491525	32.389831

Cuadro 5: Comparación desv.estandar para datos imputados

		Desv. estandar datos imputados				
	Desv. est observada	mean	median	hotdeck al	hotdeck min	hotdeck mean
Semana 1	15.93795	14.197900	14.209586	14.239867	14.690415	14.239867
Semana 2	10.71594	9.920991	9.966249	10.028131	10.012208	9.983007
Semana 3	11.64726	14.046450	14.135212	14.125388	14.159648	14.127747
Semana 4	10.45833	9.609317	9.617069	9.624814	9.618193	9.628336
Total	47.33997	44.240906	44.258915	44.650078	45.771687	44.781083

Modelos

Modelo 1

```
   Call:
glm(formula = response ~ trt_n, family = poisson(link = "log"),
    data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.060  -2.488  -1.510   0.063  18.317

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        2.1091     0.0329   64.08   <2e-16 ***
trt_ntratamiento  -0.0521     0.0460   -1.13     0.26
---
```

Modelo 2

```
Call:
glm(formula = response ~ base, family = poisson(link = "log"),
    data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.293  -1.572  -0.517   0.449  13.149

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.146604   0.037139    30.9   <2e-16 ***
base        0.021485   0.000471    45.6   <2e-16 ***
---
```

Modelo 3

```
Call:
glm(formula = response ~ age, family = poisson(link = "log"),
    data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.189  -2.300  -1.302   0.008  17.831

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.38064    0.09227   25.80  < 2e-16 ***
age         -0.01045    0.00316   -3.31  0.00094 ***
---
```

Modelo 4

Modelo 4.1

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) ['glmerMod']
Family: poisson  ( log )
Formula: response ~ logbase*trt_n + logage + (1| ID) + offset(log_time)

Fixed effects:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)                -3.610      1.075   -3.36  0.00079 ***
logbase                     0.894      0.128    7.00  2.6e-12 ***
trt_ntratamiento           -1.268      0.652   -1.94  0.05183 .
logage                      0.561      0.299    1.88  0.06077 .
logbase:trt_ntratamiento    0.302      0.197    1.54  0.12472
---
```

Modelo 4.2

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) ['glmerMod']
Family: poisson  ( log )
Formula: response ~ logbase*trt_n + logage + (1 + age| ID) +
offset(log_time)

Fixed effects:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)                -3.616      1.071   -3.37  0.00074 ***
logbase                     0.896      0.131    6.82  9.1e-12 ***
trt_ntratamiento           -1.255      0.674   -1.86  0.06263 .
logage                      0.560      0.298    1.88  0.05983 .
logbase:trt_ntratamiento    0.298      0.203    1.47  0.14123
---
```

Modelo 4.3

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) ['glmerMod']
Family: poisson  ( log )
Formula: response ~ logbase*trt_n + logage + (1 + semana| ID) +
offset(log_time)

Fixed effects:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)                -3.698      1.013   -3.65  0.00026 ***
logbase                     0.874      0.119    7.35  2.1e-13 ***
trt_ntratamiento           -1.440      0.624   -2.31  0.02092 *
logage                      0.593      0.282    2.11  0.03514 *
logbase:trt_ntratamiento    0.344      0.187    1.84  0.06549 .
```

Modelo 5

```
glm(formula = response ~ trt_n + semana + age + base + offset(log_time),
    family = poisson(link = "log"), data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.361  -1.413  -0.418   0.404  12.084

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.043164   0.124073   -0.35  0.72792
trt_ntratamiento -0.161190   0.047232   -3.41  0.00064 ***
semana2          -0.037504   0.064561   -0.58  0.56130
semana3           0.055680   0.063081    0.88  0.37741
semana4          -0.161677   0.066700   -2.42  0.01535 *
age               0.019190   0.003385    5.67  1.4e-08 ***
base              0.022816   0.000516   44.21  < 2e-16 ***
---
```

Modelo 6

Sin considerar interacción entre el tratamiento y la base

```
Call:
glm(formula = response ~ trt_n + semana4 + age + base + offset(log_time),
    family = poisson(link = "log"), data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.459  -1.409  -0.400   0.349  12.339

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.036371   0.118432   -0.31  0.75877
trt_ntratamiento -0.161190   0.047232   -3.41  0.00064 ***
semana4          -0.168471   0.055506   -3.04  0.00240 **
age               0.019190   0.003385    5.67  1.4e-08 ***
base              0.022816   0.000516   44.21  < 2e-16 ***
---
```

Considerando la interacción entre el tratamiento y la base

```
Call:
glm(formula = response ~ logbase * trt_n + logage + semana4 +
    offset(log_time), family = poisson(link = "log"), data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-5.022  -1.398  -0.299   0.797  10.988

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               -4.4959     0.3943  -11.40  < 2e-16 ***
logbase                    0.9500     0.0445   21.36  < 2e-16 ***
trt_ntratamiento          -2.0951     0.2469   -8.49  < 2e-16 ***
logage                     0.8044     0.1060    7.59  3.2e-14 ***
semana4                   -0.1685     0.0555   -3.04   0.0024 **
logbase:trt_ntratamiento   0.5535     0.0645    8.59  < 2e-16 ***
---
```

Modelo 7

Sin considerar interacción entre el tratamiento y la base

```
summary(glm_model_3, dispersion = dp)
Call:
glm(formula = response ~ trt_n + semana4 + age + base + offset(log_time),
    family = poisson(link = "log"), data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.459  -1.409  -0.400   0.349  12.339

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.03637    0.25998   -0.14   0.8887
trt_ntratamiento -0.16119    0.10368   -1.55   0.1200
semana4          -0.16847    0.12184   -1.38   0.1668
age               0.01919    0.00743    2.58   0.0098 **
base              0.02282    0.00113   20.14   <2e-16 ***
---
```

Considerando la interacción entre el tratamiento y la base

```
summary(glm_model_transform1, dispersion = dp)

Call:
glm(formula = response ~ logbase * trt_n + logage + semana4 +
    offset(log_time), family = poisson(link = "log"), data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-5.022  -1.398  -0.299   0.797  10.988

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -4.4959     0.8519   -5.28  1.3e-07 ***
logbase                    0.9500     0.0961    9.89  < 2e-16 ***
trt_ntratamiento          -2.0951     0.5333   -3.93  8.6e-05 ***
logage                     0.8044     0.2289    3.51  0.00044 ***
semana4                   -0.1685     0.1199   -1.41  0.16002
logbase:trt_ntratamiento   0.5535     0.1393    3.97  7.1e-05 ***
---
```

Modelo 8

```
Call:
glm(formula = response ~ logbase * trt_n + semana4 + logage +
    offset(log_time), family = quasipoisson, data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-5.022  -1.398  -0.299   0.797  10.988

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -4.4959     0.8519   -5.28  3.0e-07 ***
logbase                    0.9500     0.0961    9.89  < 2e-16 ***
trt_ntratamiento          -2.0951     0.5333   -3.93  0.00011 ***
semana4                   -0.1685     0.1199   -1.41  0.16137
logage                     0.8044     0.2289    3.51  0.00053 ***
logbase:trt_ntratamiento   0.5535     0.1393    3.97  9.4e-05 ***
---
```

Modelo 9

```
Call:
MASS::glm.nb(formula = response ~ logbase * trt_n + logage +
    semana4 + offset(log_time), data = data, init.theta = 2.737032419,
    link = log)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-3.069  -0.804  -0.208   0.377   3.760

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                -3.6413     0.7552   -4.82  1.4e-06 ***
logbase                     0.8946     0.0888   10.08  < 2e-16 ***
trt_ntratamiento           -1.3065     0.4684   -2.79   0.0053 **
logage                      0.6073     0.2079    2.92   0.0035 **
semana4                    -0.1660     0.1141   -1.45   0.1457
logbase:trt_ntratamiento    0.3290     0.1382    2.38   0.0173 *
```

Modelo 10

```
Call:
glm(formula = response ~ logbase * trt_n + logage + offset(log_time),
    family = quasipoisson(link = "log"), data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.924  -1.394  -0.227   0.801  11.205

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                -4.5354     0.8609   -5.27  3.1e-07 ***
logbase                     0.9500     0.0971    9.78  < 2e-16 ***
trt_ntratamiento           -2.0951     0.5392   -3.89  0.00013 ***
logage                      0.8044     0.2314    3.48  0.00061 ***
logbase:trt_ntratamiento    0.5535     0.1408    3.93  0.00011 ***
---
```

Problema 2

Analisis exploratorio

Test Log-Rank

Sexo:

```
Call:
survdiff(formula = datosSurv ~ sexo, data = surv_hd)

          N Observed Expected (O-E)^2/E (O-E)^2/V
sexo=F 3623      110      106     0.153     0.569
sexo=M 1332       35       39     0.416     0.569

 Chisq= 0.6  on 1 degrees of freedom, p= 0.5
```

Comorbilidades:

```
Call:
survdiff(formula = datosSurv ~ comorbilidades, data = surv_hd)

                    N Observed Expected (O-E)^2/E (O-E)^2/V
comorbilidades=0 4944      136  144.807     0.536       402
comorbilidades=1   11        9    0.193   401.838       402

 Chisq= 402  on 1 degrees of freedom, p= <2e-16
```

Edad:

```
Call:
survdiff(formula = datosSurv ~ age_group, data = surv_hd)

                   N Observed Expected (O-E)^2/E (O-E)^2/V
age_group=21-30  796       27     23.2     0.614     0.731
age_group=30-40 1728       57     50.5     0.842     1.292
age_group=40-50 1207       37     35.3     0.081     0.107
age_group=+50   1224       24     36.0     3.993     5.312

 Chisq= 5.5  on 3 degrees of freedom, p= 0.1
```