Taller de Análisis de datos - Problema 0

Jésica Charaf e Ignacio Spiousas

27 de octubre de 2023

Problema 0

Se dan a continuación los tiempos de sobrevida (en unidades de 10 horas) de animales, sometidos a 3 tipos de veneno, y 4 tratamientos antitóxicos. Cada combinación veneno-tratamiento se prueba con 4 animales. Describir la influencia de ambos factores en la sobrevida. ¿Hay algún tratamiento demostrablemente mejor?.

	Tratamientos			
Veneno	A	B	C	D
I	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
II	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
III	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

Resolución

Los datos contienen 48 observaciones con las siguientes variables involucradas:

- Sobrevida: Tiempo de sobrevida del animal (en unidades de 10 horas)
- Tratamiento: Tipo de tratamiento antitóxicos (A, B, C y D)
- Veneno: Tipo de veneno (I, II y III)

Primero, realizamos algunos gráficos exploratorios para tener una visualización inicial de los datos.

En la figura 1 se muestran el tiempo de sobrevida por tipo de veneno y tratamiento. Los puntos translucidos corresponden a los datos individuales mientras que el punto lleno corresponde a la media muestral. Allí podemos observar que el tiempo medio de sobrevida en el tratamiento B resulta, en general, mayor que en el resto. Sin embargo, los tiempos de sobrevida de todos los tratamientos parecen depender fuertemente del tipo de veneno.

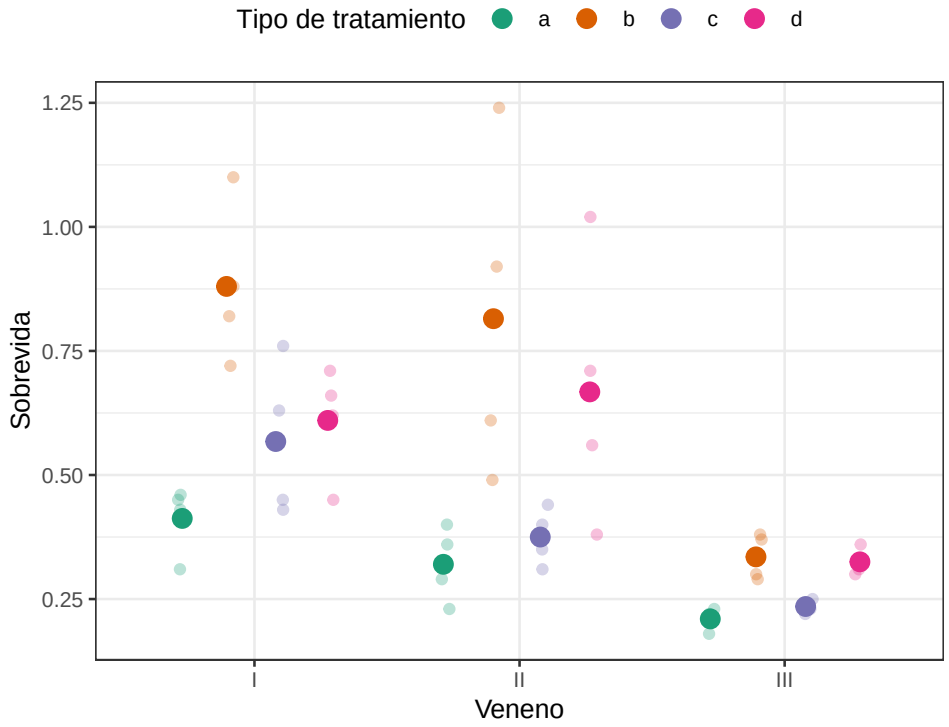

Figure 1: Tiempos de sobrevida en función del tipo de tratamiento para cada tipo de veneno. En puntos translucidos se muestran los puntos individuales y en puntos llenos las medias muestrales.

A continuación realizamos un gráfico para analizar la interacción entre el tratamiento y el tipo de veneno (figura 2), donde se ven representados los valores medios de la variable sobrevida para cada tratamiento en función del tipo de veneno.

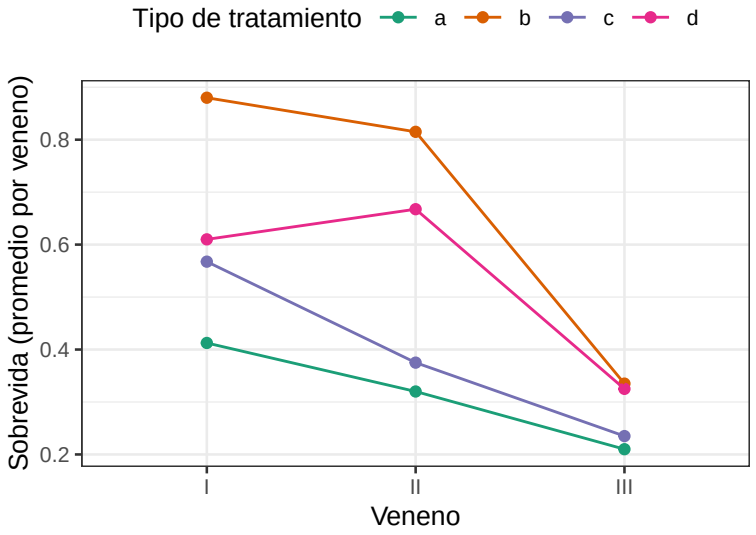

Figure 2: Gráfico de interacción entre tratamiento y tipo de veneno. El tipo de veneno está representado en el eje x mientras que el tipo de tratamiento está codificado por el color de la línea.

Las curvas resultantes no se ven paralelas, por lo que no es posible deducir que no haya ninguna interación entre las variables veneno y tratamiento. Sin embargo, se puede apreciar que el tratamiento B está por encima del resto, siguiendo luego en orden los tratamientos D, C y A, independientemente del tipo de veneno. A su vez, se observa que el veneno III es el que más disminuye el tiempo de sobrevida independientemente del tratamiento y el veneno I pareciera ser el que menos lo disminuye en la mayoría de los tratamientos.

Modelo aditivo vs. modelo con interacción

Primero ajustamos el modelo regresión lineal aditivo:

```
sobrevida ~ veneno + tratamiento (mod1)
```

Para explorar si se verifican los supuestos del modelo, utilizamos la función `check_model()` del paquete *{performance}* y graficamos un QQ-plot, los VIF y la relación de los residuos con los valores predichos (figura 3).

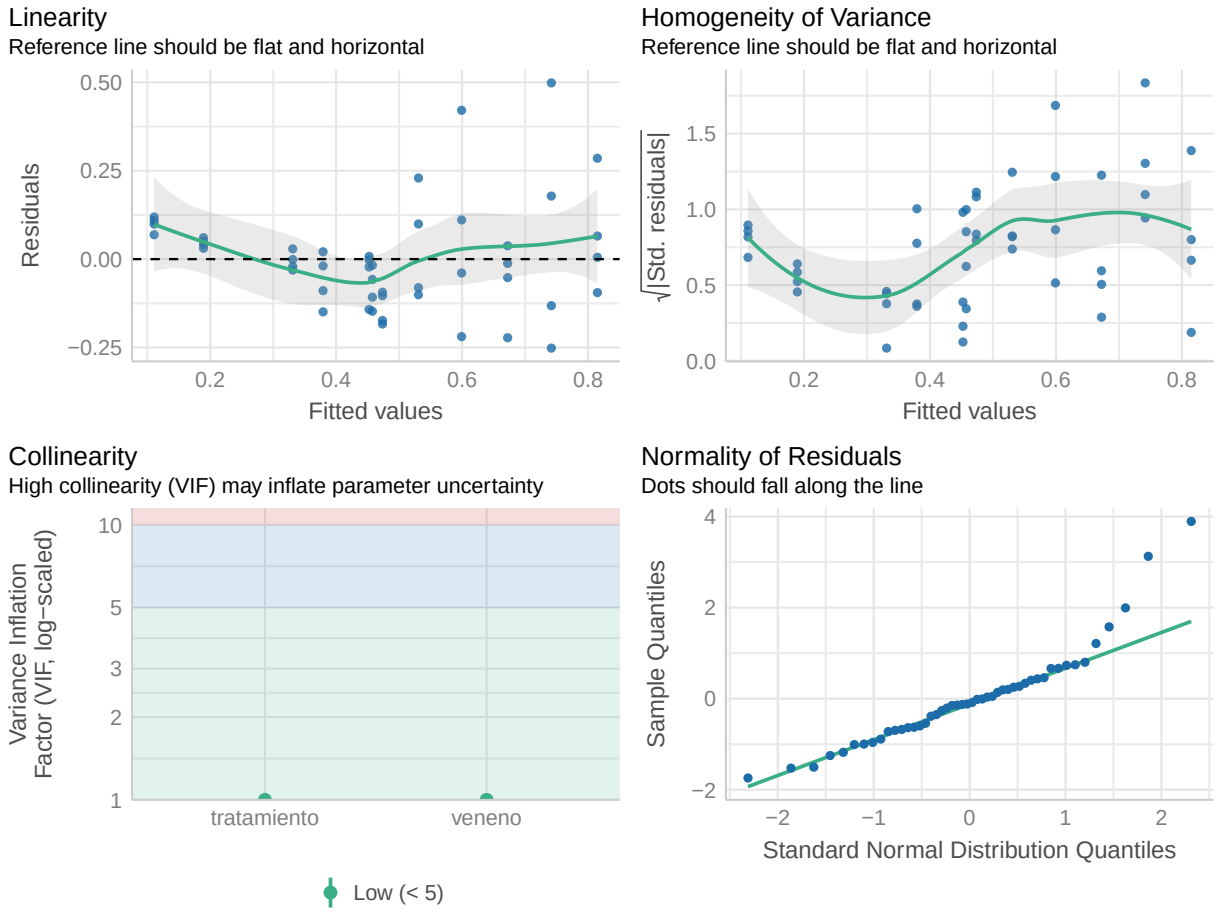

Figure 3: Gráficas de diagnóstico del modelo aditivo.

Por una parte, podemos observar cierta estructura en el gráfico de los residuos contra los valores predichos, con lo cual no pareciera cumplirse el supuesto de homoscedasticidad. Además, si miramos el QQ-plot de los residuos encontramos varias observaciones no alineadas lo que indicaría que tampoco se cumple el supuesto de normalidad.

Si ajustamos el modelo de regresión lineal con interacción

`sobrevida ~ veneno + tratamiento + veneno : tratamiento (mod 2)`

y realizamos los mismos gráficos (figura 4), obtenemos algo similar en relación a los supuestos de normalidad y homoscedasticidad. Además, en el gráfico de los VIF podemos ver que superan el valor de 10, indicando que hay colinealidad entre las variables.

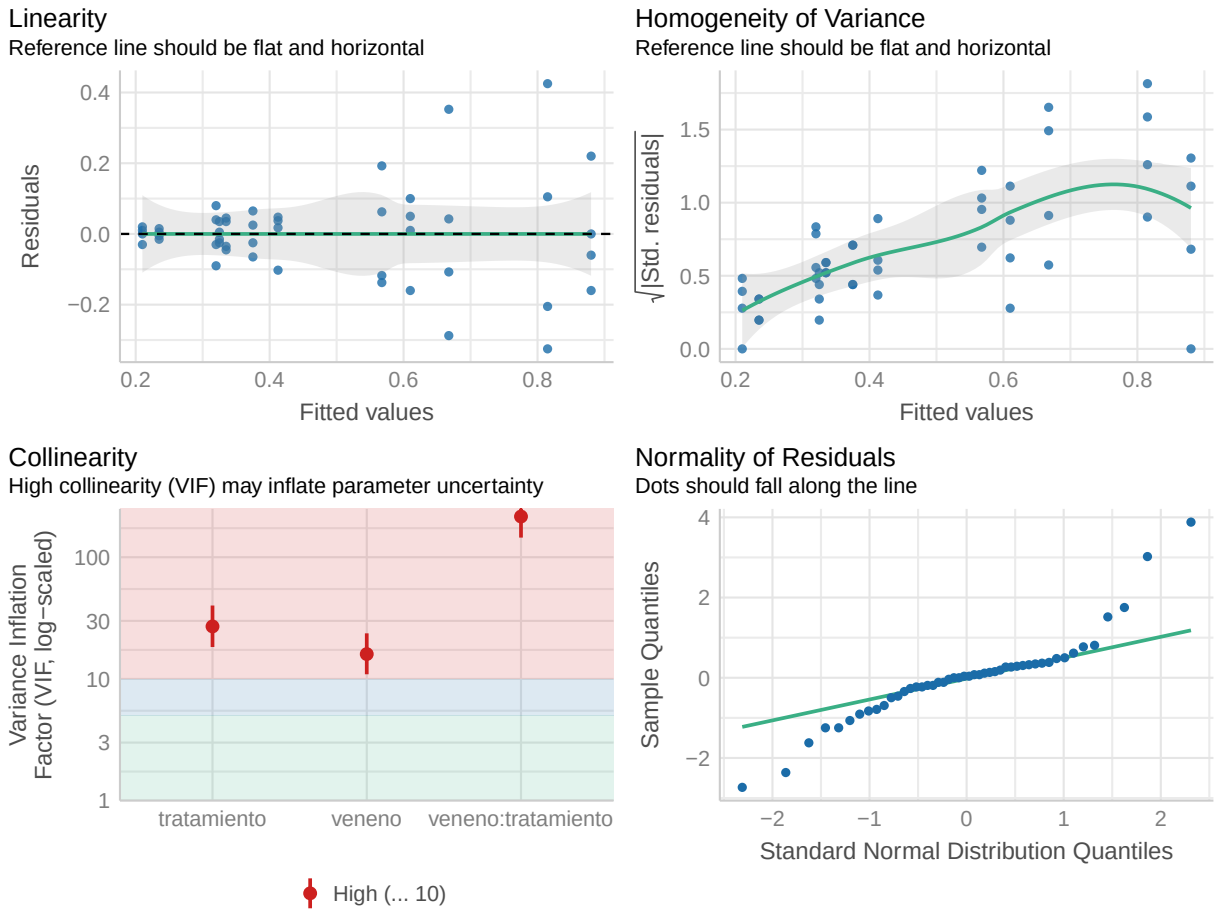

Figure 4: Gráficas de diagnóstico del modelo con interacción.

Realizamos un test ANOVA para comparar el modelo 1 con el modelo 2 y analizar si la interacción resulta significativa. La hipótesis nula consiste en que los coeficientes de la interacción son todos iguales a cero y el p-valor obtenido es 0.112. De esta manera, a partir de este test no habría evidencia suficiente para decir que la interacción es significativa (tabla 1).

Table 1: Tabla de anova comparando el modelo aditivo con el modelo con interacción.

term	df.residual	rss	df	sumsq	statistic	p.value
sobrevida ~ veneno + tratamiento	42	1.050863	NA	NA	NA	NA
sobrevida ~ veneno * tratamiento	36	0.800725	6	0.2501375	1.874333	0.1122506

Teniendo en cuenta el análisis previo, optamos por desestimar la interacción ya que no tenemos evidencia clara de que el aporte sea significativo, incorporarla no mejora los supuestos del modelo lineal y consideramos conveniente trabajar con un modelo más sencillo para intepretar.

Para estudiar si alguno de los tratamientos es mejor que el resto vamos a comparar de a pares. Para esto es necesario utilizar alguna estrategia que tenga en cuenta la inflación del error de tipo I debido a las comparaciones múltiples. En este caso utilizamos el test de comparaciones múltiples de medias de Tukey (nivel = 0.05). Los resultados obtenidos se representan como intervalos de confianza en la figura 5.

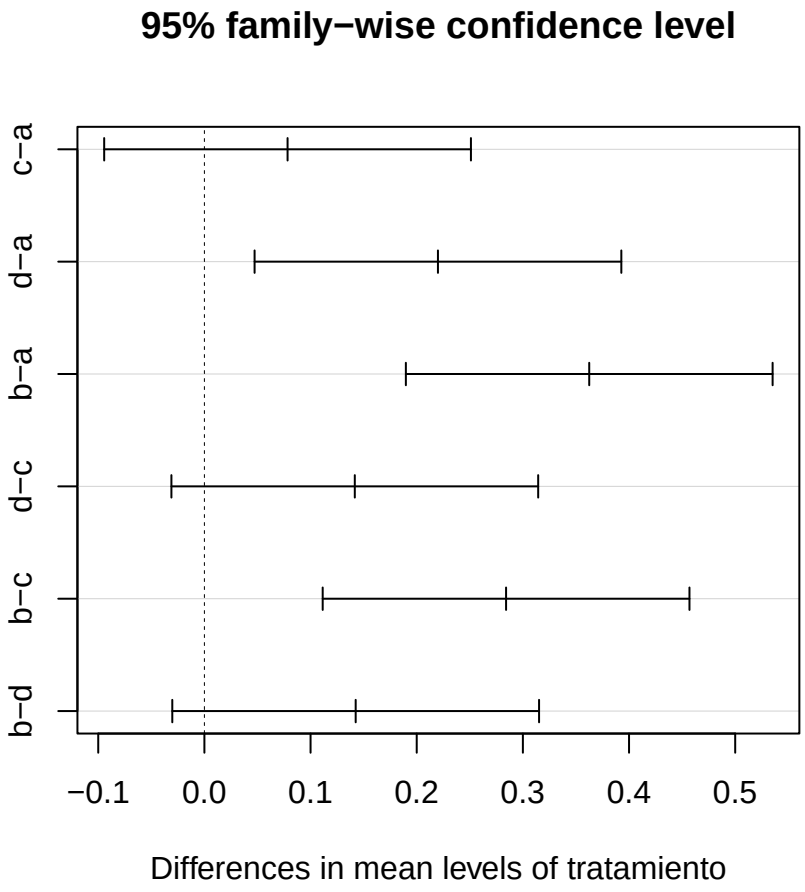

Figure 5: Intervalos de confianza, corregidos por Tukey, del modelo aditivo.

En la figura 7 vemos que el tratamiento B resulta superior que los tratamientos A y C y no hay evidencia para decir que se distingue del tratamiento D, dado que el 0 pertenece al intervalo de confianza de la comparación b-d. A su vez, observamos que el tratamiento D resulta mejor que el A pero no se puede distinguir del tratamiento C. Y por último, no tenemos evidencia para decir que los tratamientos C y A se diferencian.

Transformamos los datos

Con el objetivo de mejorar la adecuación de los datos a los supuestos del modelo de regresión lineal, probamos incorporar distintas transformaciones a la variable sobrevida.

Para empezar, ajustamos el modelo:

```
log(sobrevida) ~ veneno + tratamiento (mod3)
```

Volvemos a generar las gráficas de diagnóstico del modelo (figura 6) y no vemos mejoras significativas en relación a los supuestos de homoscedasticidad y normalidad.

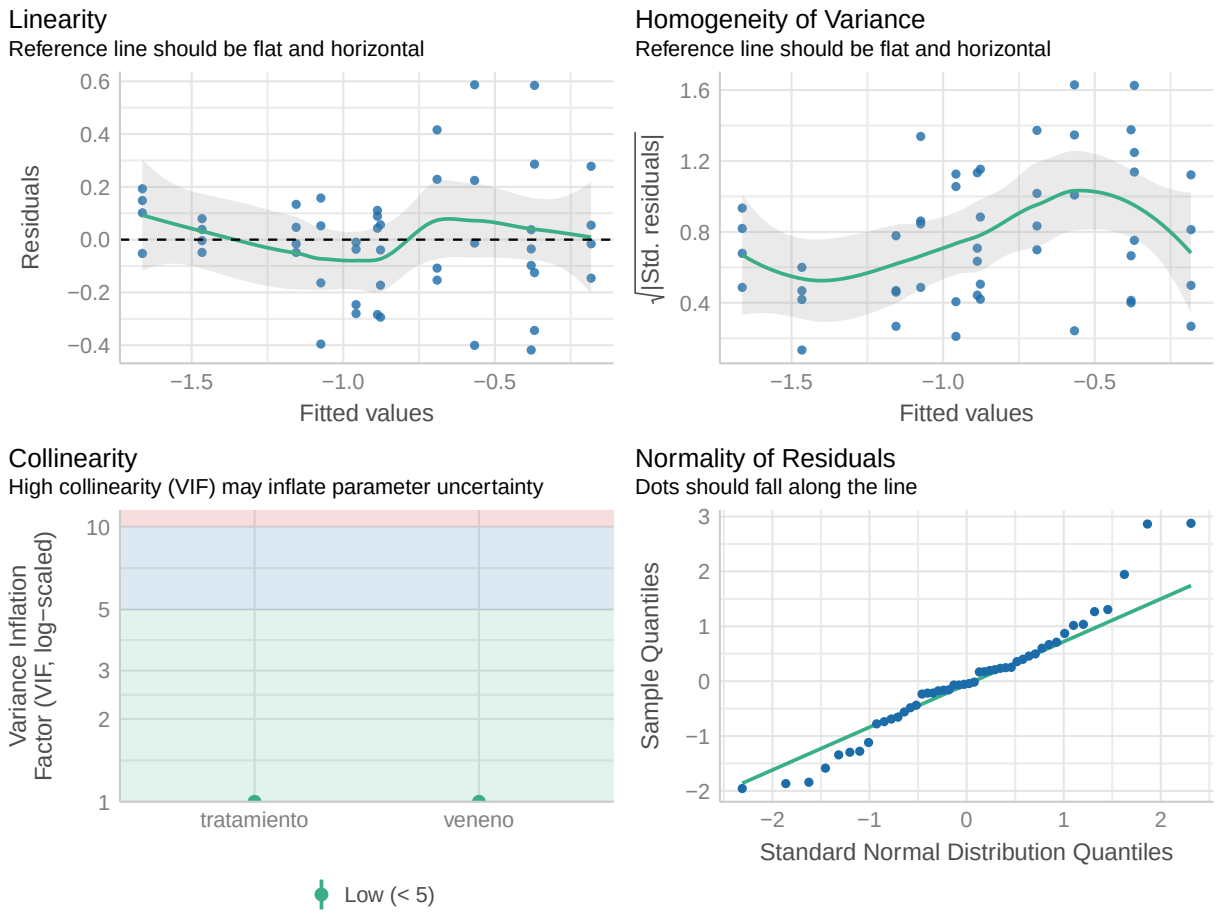

Figure 6: Gráficas de diagnóstico del modelo con la sobrevida transformada con log()

Por lo tanto, optamos por probar con una transformación de la familia Box-Cox:

$$f(x) = \frac{x^\lambda - 1}{\lambda}, \lambda \neq 0$$

Con el comando *boxcox* de R calculamos el valor óptimo para λ, obteniendo $\lambda = -0.75$. Luego, tomando este valor ajustamos el modelo:

```
f(sobrevida) ~ veneno + tratamiento (mod 4)
```

A partir del gráficos de los residuos vs. valores predichos y del QQ-plot (figura 7) podemos ver que mejoran bastante los supuestos de homoscedastidad y de normalidad.

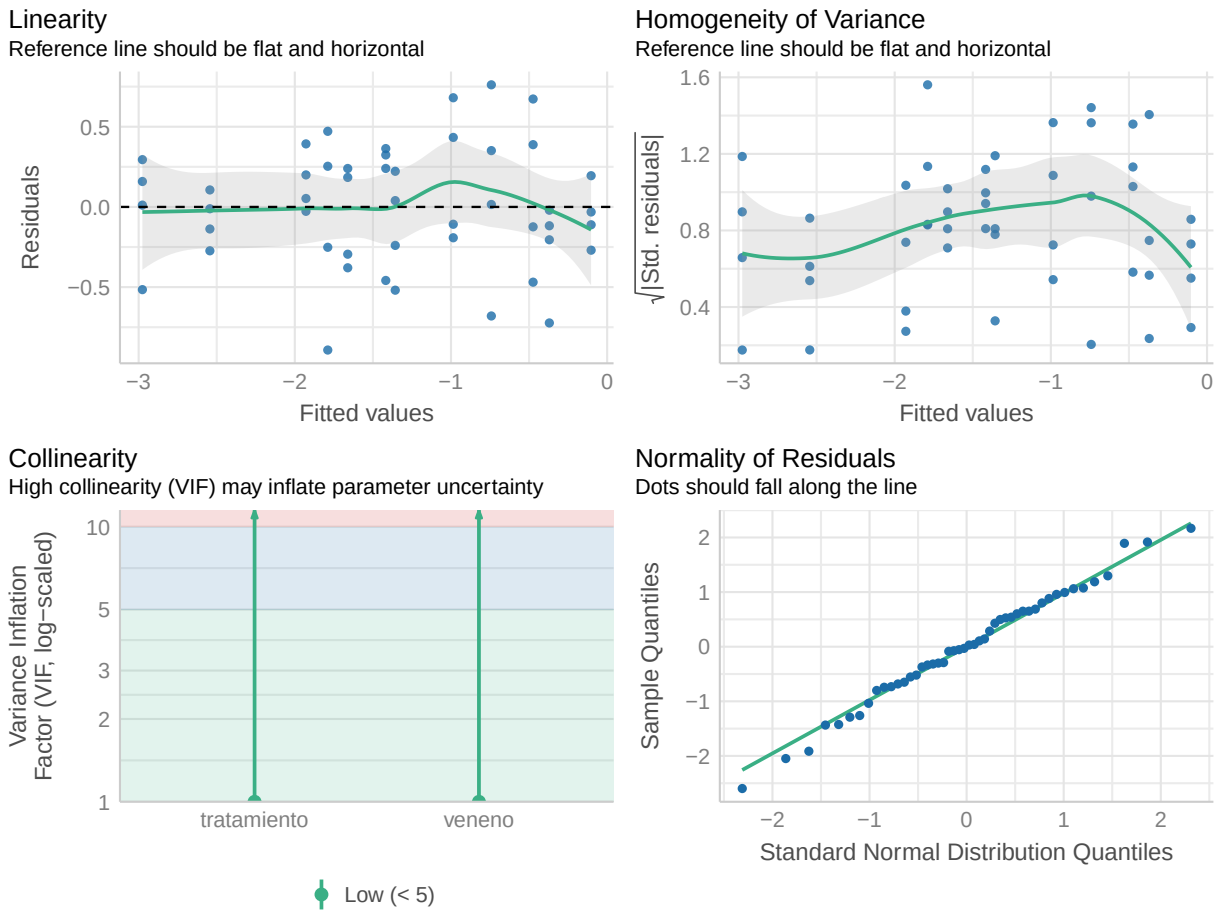

Figure 7: Gráficas de diagnóstico del modelo con la sobrevida transformada con la función boxcox ($\lambda = -0.75$)

También realizamos gráficos de interacción entre las variables tratamiento y veneno, representando los valores medios de la transformación realizada a la variable sobrevida (figura 8).

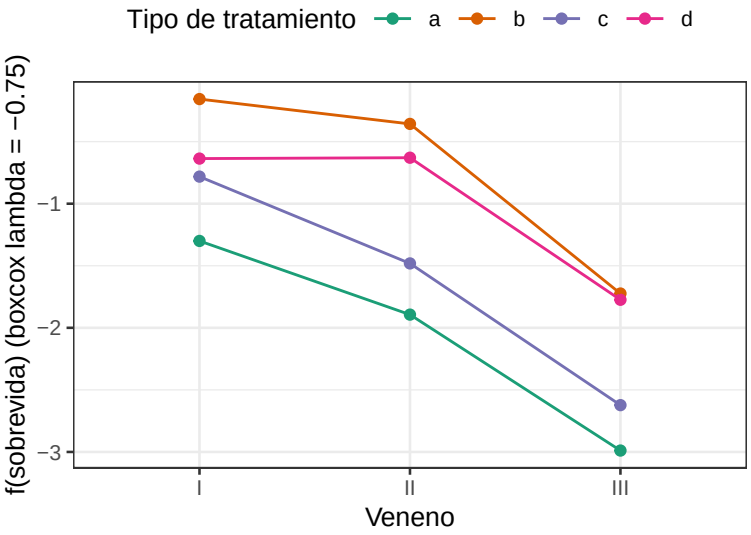

Figure 8: Gráfico de interacción entre tratamiento y tipo de veneno en la sobrevida luego de aplicarle la función boxcox ($\lambda = -0.75$). El tipo de veneno está representado en el eje x mientras que el tipo de tratamiento está codificado por el color de la línea.

Si bien las curvas no resultan paralelas, vemos que mejora un poco esta característica y que el orden de los tratamientos se mantiene independientemente del tipo de veneno. Lo mismo ocurre para cada veneno en función del tratamiento y, además, ya no se observa entrecruzamiento.

Analizando el summary del ajuste del modelo 4 (tabla 2), observamos que todos los coeficientes resultan significativos. Se puede destacar que el coeficiente correspondiente al veneno III es -1.558, el mayor en valor absoluto, corroborando nuestra observación inicial de que el veneno III es el que mayor impacto tiene en el tiempo de sobrevida. A la vez, el coeficiente correspondiente al tratamiento B tiene un valor de 1.315 y es el mayor de todos los coeficientes de los tratamientos, lo que indica un aporte superior en el tiempo medio sobrevida.

Table 2: Estimadores para los coeficientes del modelo con la transformación boxcox (lambda = -0.75).

Predictor	B	SE	t	p
(Intercept)	-1.418	0.138	-10.24	<0.001
venenoII	-0.371	0.138	-2.68	0.010
venenoIII	-1.558	0.138	-11.26	<0.001
tratamientob	1.315	0.160	8.23	<0.001
tratamientoc	0.432	0.160	2.70	0.010
tratamientod	1.047	0.160	6.55	<0.001

A partir de un test ANOVA evaluamos si las variables veneno y tratamiento resultan significativas obteniendo los p-valores $5.161e-14$ y $4.901e-10$ respectivamente, con lo cual vemos que ambas resultan relevantes para explicar el tiempo de sobrevida.

A su vez, consideramos el modelo con interacción

```
f(sobrevida) ~ veneno + tratamiento + veneno : tratamiento
```

y realizamos un test ANOVA para evaluar si la interacción resulta significativa. Obtuvimos un p-valor igual a 0.486 (tabla 3), con lo cual no hay evidencia para considerar que es relevante la interacción en este modelo.

Table 3: Tabla de anova del modelo con la transformación boxcox e interacción.

term	df.residual	rss	df	sumsq	statistic	p.value
boxcox_sobrevida ~ veneno * tratamiento	36	5.575	NA	NA	NA	NA
boxcox_sobrevida ~ veneno + tratamiento	42	6.439	-6	-0.864	0.93	0.486

Para concluir, a partir del ajuste del modelo 4 realizamos un test de comparaciones múltiples de medias de Tukey de nivel 0.05 para comparar los tratamientos y analizar si alguno resulta mejor que los demás. Los resultados obtenidos, en forma de intervalos de confianza, están representados en la figura 9.

Los resultados nos permiten concluir, al igual que con el modelo 1, que el tratamiento B resulta mejor que el A y el C, mientras que no hay evidencia de que se diferencie con el tratamiento D. Por otra parte, en este caso también se concluye que el tratamiento de D resulta superior al A y C. Además el tratamiento C parece ser mejor que el A.

Si bien el test de Tukey de nivel $\alpha = 0.05$ no nos permite distinguir entre el tratamiento B y D, en base a todo el análisis realizado consideramos que el tratamiento B es el más conveniente.

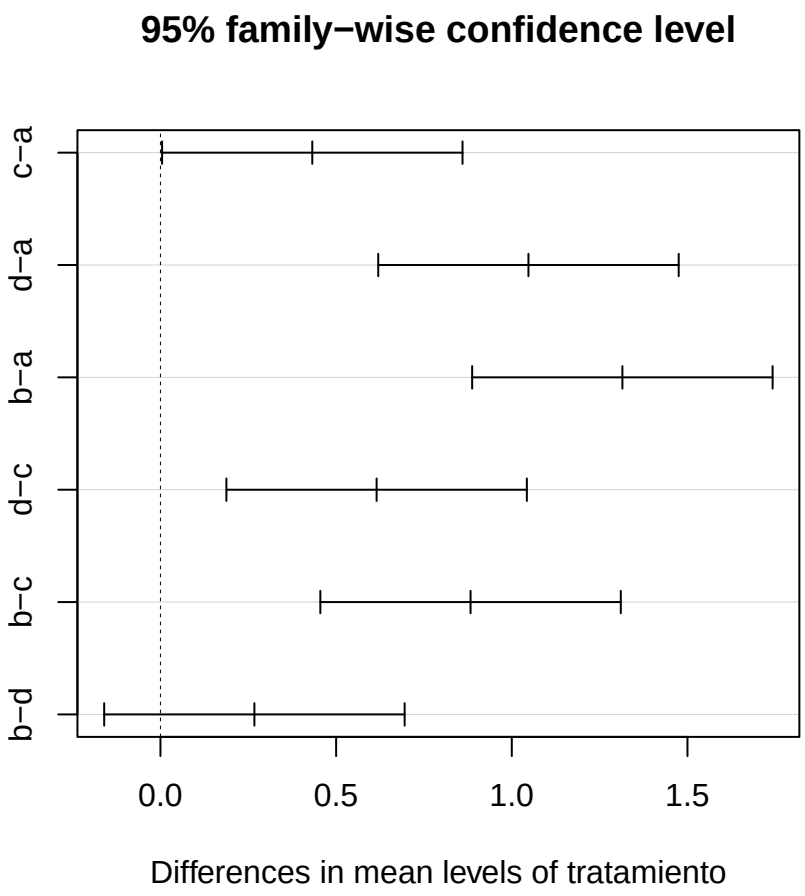

Figure 9: Intervalos de confianza, corregidos por Tukey, del modelo con la sobrevida tranformada con la función boxcox ($\lambda = -0.75$)