5. Problema ANOVA. Medicamentos.

I. Análisis descriptivo y/o visualización de datos.

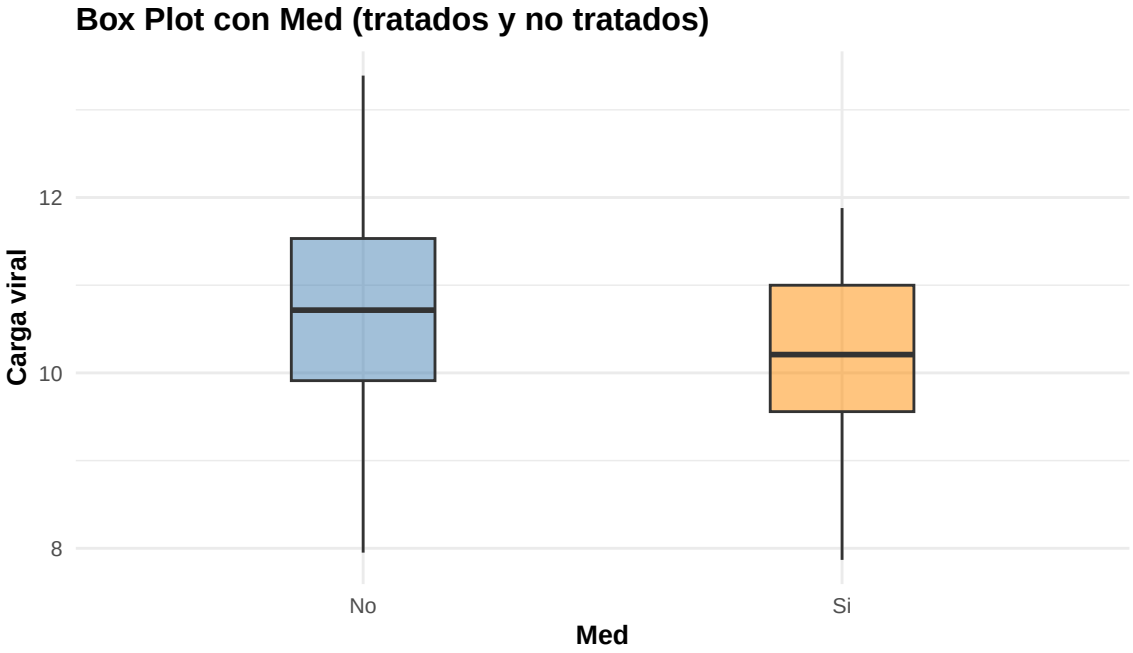
En la base de datos `Ejercicio5B` se tiene información del índice de carga viral (Y) y si se aplicó o no el medicamento contra Covid (Med) para un grupo de 100 personas. A 50 personas se le aplicó el medicamento.

En el siguiente Cuadro se muestra la estadística descriptiva del dato numérico, que es el índice de carga viral, es posible observar los pacientes presentaron un mínimo de 7.8681037 y un maximo de 13.3899626, con una media de 10.4800674. La desviación estándar de 1.1477774 es pequeña, por lo que parece que los datos no son tan dispersos.

Table 1:

Statistic	N	Mean	St. Dev.	Min	Max
datos$Y	100	10.480	1.148	7.868	13.390

En la siguiente Gráfica de caja y bigotes (brazos), podemos observar que hay una mediana mayor de carga viral para los individuos no tratados, con respecto a los tratados, además de una mayor dispersión de los datos para los no tratados. No se observaron outliers o valores atípicos.



II. Planteamiento y estimación del modelo.

Para ver si la menor carga viral está asociada con la aplicación del medicamento planteamos un modelo de regresión donde la variable dependiente es la carga viral y_i y la variable independiente x_i se puede ver como categórica, donde $x_i = 1$ si el paciente es tratado y $x_i = 0$ si no.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Verificaremos que el modelo presente linealidad, además de homocedasticidad y no autocorrelación en los errores, para poder interpretar los resultados del ajuste del modelo que se muestra en la columna (1) del Cuadro al final de esta sección.

III. Validación de supuestos y pruebas de hipótesis.

Haremos una prueba de hipótesis, para responder a la pregunta de si existe una relación entre la aplicación del medicamento y la disminución de la carga viral. Es decir, planteamos la hipótesis nula $H =: \beta_1 > 0$ y la alternativa $H_a : \beta_1 < 0$. En la prueba `Simultaneous Tests for General Linear Hypotheses` se rechaza H_0 con un nivel de confianza del 95%, pues el p-value asociado es de 0.0204 y t-value de -2.074. Sin embargo, para que este resultado tenga validez y para que también las pruebas de hipótesis individuales y global del modelo de la columna (1) del cuadro al final de esta sección tengan validez y podamos interpretar los coeficientes estimados, debemos de hacer las pruebas de cumplimiento de los supuestos del modelo de regresión lineal.

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = Y ~ Med, data = datos)
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(<t)
## 1 >= 0  -0.4682     0.2258  -2.074 0.0204 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Simultaneous Tests for General Linear Hypotheses				
Fit: lm(formula = Y ∼Med, data = datos)				
Linear Hypotheses:				
	Estimate	Std. Error	t value	Pr(<t)
1 >= 0	-0.4682	0.2258	-2.074	0.0204*
Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			
(Adjusted p values reported – single-step method)				

En el siguiente Cuadro se pueden observar las pruebas de Shapiro-Wilk, Breusch-Pagan y Durbin-Watson para el Modelo 1, que plantean la hipótesis nulas de normalidad, homoscedasticidad y no autocorrelación, respectivamente. Se concluye que el Modelo 1 presenta normalidad de los errores, sin embargo presenta autocorrelación y heteroscedasticidad. Por lo que tendríamos que hacer algunos ajustes al modelo, con algunos tratamientos a las variables, si quisiéramos usarlo para inferencia.

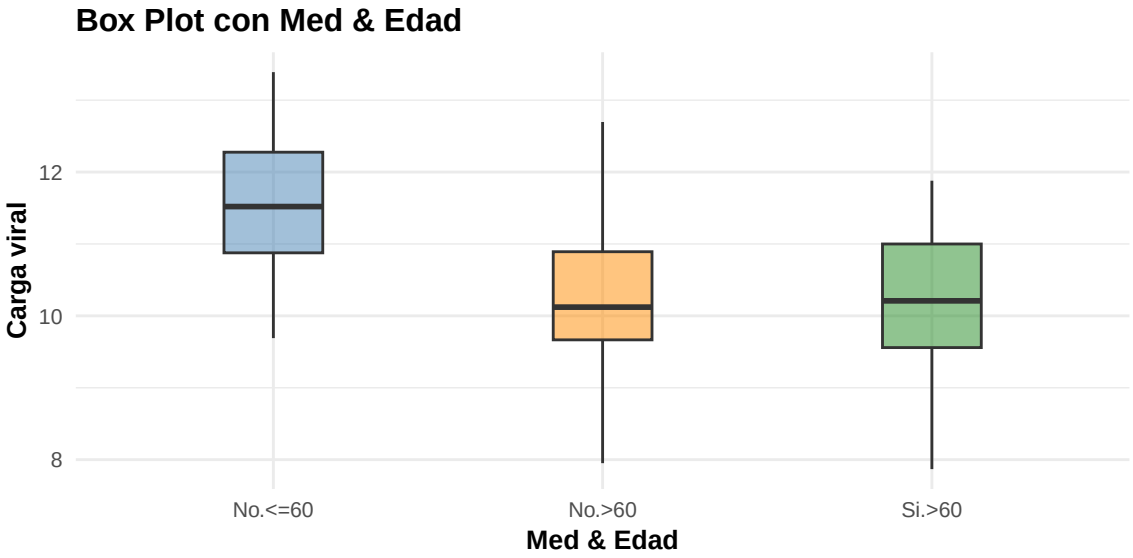
	1
Normality (Shapiro-Wilk)	0.532
Homoscedasticity (Breusch-Pagan)	0.048
Autocorrelation of residuals (Durbin-Watson)	0.021

IV. Consideración de la variable Edad.

Tenemos un total de 100 pacientes, de los cuales 80 tienen más de 60 años y a la mitad de todos los pacientes se le aplicó el medicamento (tratados). En la base de datos tenemos 20 observaciones de personas no tratadas menores de 60 años y no tenemos personas de ese grupo de edades que sean tratadas por el medicamento, lo que podría sesgar los resultados.

A continuación podemos observar en la gráfica de caja y bigotes que a los pacientes menores de 60 años a quienes no se les aplicó la vacuna (no fueron tratadas) tienen una mayor carga viral, sin embargo no disponemos de datos para personas menores a 60 años a quienes se les aplicó el medicamento, para una mejor comparación. Por otro lado, para mayores de 60 años, parece no haber una diferencia clara entre los pacientes a los que se le aplicó la vacuna y a las personas a las que no se le aplicó, pues ambos grupos tienen cargas virales menores pero muy parecidas. Finalmente, no se detectaron datos atípicos (outliers) en los datos. Por

lo tanto, consideramos que los rsultados anteriores no son contundentes, por lo que convendría controlar por la variable de Edad.



V. Planteamiento del nuevo modelo y pruebas de hipótesis.

En este caso planteamos el mismo modelo, pero como no tenemos individuos tratados y no tratados menores de 60 años en la base de datos (solamente tenemos a los no tratados), tomaremos en cuenta a los grupos homogéneos de tratados y no tratados mayores o iguales a 60 años, por lo que nuestra base de datos se reduce de 100 a 80 observaciones.

En la columna (2) del Cuadro a final de esta sección se muestran los resultados correspondientes a la regresión lineal para los pacientes tratados y no tratados mayores a 60 años. La prueba global F muestra que no es posible rechazar la hipótesis nula de que los coeficientes asociados a las variables explicativas son cero, pues el p-value asociado es grande, incluso mayor a 0.1. Al parecer, las conclusiones obtenidas al tomar toda la muestra estaban sesgadas por el grupo de edad, pues cuando quitamos a los no tratados menores de 60 años que tenían una carga viral bastante importante, los resultados cambiaron. Como en la prueba global no es posible rechazar la hipítesis nula, no es posible continuar el análisis con este modelo, y esto está relacionado con el no rechazo de la hipótesis nula de la prueba individual t-student para β_1, es decir, que la variable explicativa no es estadísticamente significativa en el modelo de regresión lineal.

NOTA: Si tomáramos todos los datos, sin considerar los grupos heterogéneos, para el coeficiente estimado asociado a la aplicación o no al medicamento, no se puede rechazar la hipótesis nula de $\beta_1 = 0$ contra la alternativa de $\beta_1 \neq 0$. Esto se muestra en la columna (3) del cuadro al final de esta sección. Además, la variable estadísticamente significativa es la edad, la prueba global F también rechaza H_0 de que todos los coeficientes estimados son cero. Adicionalmente, para este modelo se cumplen los tres supuestos más importantes, como la normalidad, no autocorrelación y homocedasticidad, como se muestra en el Cuadro correspondiente a las pruebas Shapiro-Wilk, Breusch-Pagan y Durbin-Watson. Sin embargo, no podemos continuar por este camino, en primera porque estamos analizando datos heterogéneos, tal vez si tuviéramos a los individuos menores de 60 años con tratamiento, podríamos analizarlo, y en segundo lugar, la respuesta a que si hay un efecto del medicamento parecería ser negativa y solamente dependería de la variable edad.

	1
Normality (Shapiro-Wilk)	0.818
Homoscedasticity (Breusch-Pagan)	0.253
Autocorrelation of residuals (Durbin-Watson)	0.206

Table 4:

	Dependent variable:		
	Y		
	(1)	(2)	(3)
MedSi	−0.468**	0.029	0.029
	(0.226)	(0.245)	(0.242)
Edad>60			−1.244***
			(0.302)
Constant	10.714***	10.217***	11.460***
	(0.160)	(0.194)	(0.234)
Observations	100	80	100
R^2	0.042	0.0002	0.184
Adjusted R^2	0.032	−0.013	0.168
Residual Std. Error	1.129 (df = 98)	1.062 (df = 78)	1.047 (df = 97)
F Statistic	4.299** (df = 1; 98)	0.014 (df = 1; 78)	10.960*** (df = 2; 97)
Note:			*p<0.1; **p<0.05; ***p<0.01