Ejercicio 1

Equipo

2024-03-26

1. Regresión lineal múltiple.

i) Modelo de RLM reducido para E(bpsystol; bmi, sex, age) con datos originales.

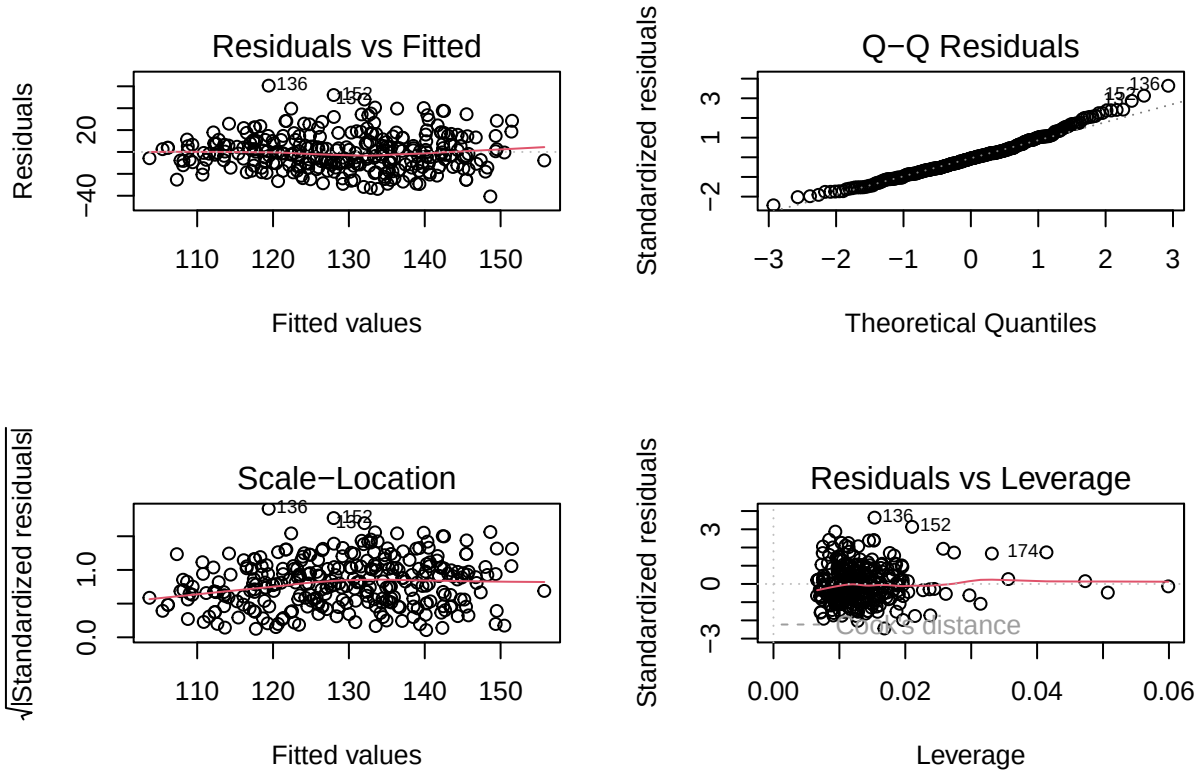
Para analizar si existe una asociación entre la presión arterial sistólica (bpsystol) como variable dependiente y el índice de masa corporal (bmi), ajustaremos un modelo de regresión lineal múltiple, considerando el sexo (sex: 1-hombre y 2-mujer con nivel de referencia hombre) y la edad (age) de los pacientes. Para ello usaremos la base de datos `reg1B.csv` con 295 pacientes, 142 hombres y 153 mujeres, de entre 20 y 74 años. En el siguiente Cuadro se muestran los resultados del modelo planteado, sin pretratamiento de los datos.

La prueba global F muestra un p-value menor a 0.05, por lo que rechazamos la hipótesis nula de que los parámetros estimados son cero, es decir, podemos decir que al menos un coeficiente estimado es distinto de cero, por lo que el modelo es estadísticamente significativo al nivel de confianza del 95%. Las pruebas individuales también rechazan la hipótesis nula con la preba $t - student$, es decir, todos los coeficientes son significativos al 5% de significancia estadística, pues re rechaza la hipótesis nula de que son en lo individual cero.

Para poder tener una interpretación válida de los coeficientes, veremos si el modelo cumple con los supuestos del modelo de regresión lineal. La Gráfica **Residuals vs Fitted**, se utiliza para comprobar los supuestos de relación lineal, una línea horizontal, sin patrones distintos, es indicación de una relación lineal, lo que es bueno en nuestro caso. La Gráfica **Normal Q-Q Residuals**, se utiliza para examinar si los residuos se distribuyen normalmente, es bueno que los puntos residuales sigan la línea recta discontinua, en nuestro caso parece que no todo se ajusta bien, pues tenemos muchos vaores que no siguen la linea. La Gráfica **Scale-Location**, se utiliza para comprobar la homogeneidad de la varianza de los residuos (homoscedasticidad), la línea horizontal con puntos igualmente distribuidos es una buena indicación de homocedasticidad, este es el caso en nuestro ejemplo, donde no tenemos un problema de heterocedasticidad. La Gráfica **Residuals vs Leverage**, se utiliza para identificar casos de valores influyentes, es decir, valores extremos que podrían influir en los resultados de la regresión cuando se incluyen o excluyen del análisis, al parecer ningún valor sale de la distancia de Cook.

Table 1:

	Dependent variable:
	bpsystol
bmi	1.208***
	s.e: (0.202)
	t value: 5.995
	Pr(>|t|): 6.02e-09
sex2	−5.664***
	s.e: (1.964)
	t value: -2.884
	Pr(>|t|): 0.00421
age	0.484***
	s.e: (0.059)
	t value: 8.264
	Pr(>|t|): 5.03e-15
Constant	84.160***
	s.e: (6.037)
	t value: 13.942
	Pr(>|t|): < 2e-16
Observations	295
R^2	0.310
Adjusted R^2	0.302
Residual Std. Error	16.784 (df = 291)
F Statistic	43.497*** (df = 3; 291); p-value: < 2.2e-16
Note:	*p<0.1; **p<0.05; ***p<0.01



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

	1
Normality (Shapiro-Wilk)	0.001
Homoscedasticity (Breusch-Pagan)	0.095
Autocorrelation of residuals (Durbin-Watson)	0.981

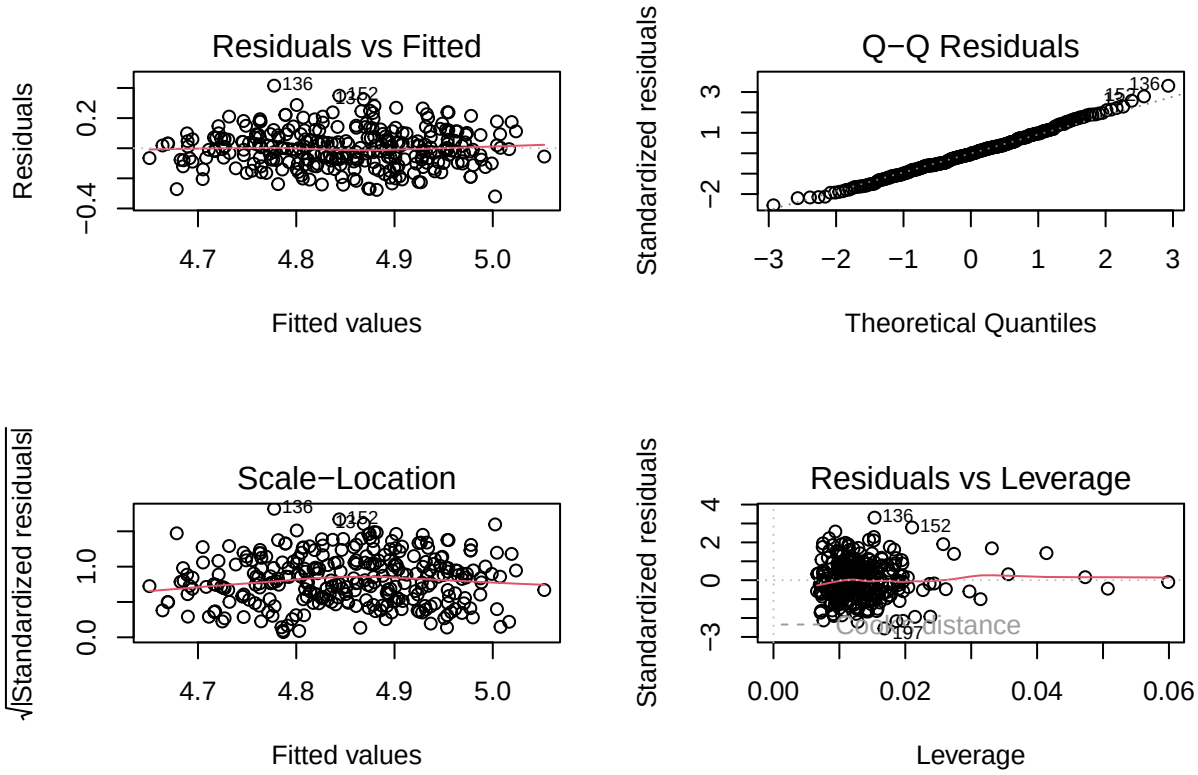
ii) Modelo adecuado con transformación de datos.

Por simplicidad no consideraremos en el modelo interacciones entre las variables y se propone una transformación logarítmica de la variable dependiente. Para este modelo, se observa en el siguiente Cuadro que la prueba global F rechaza la hipótesis nula, y las pruebas $t - student$ individuales de los coeficientes estimados también rechaza las hipótesis nulas de manera individual e independiente.

En las siguientes gráficas podemos observar en **Residuals vs Fitted** que se conserva la linealidad, en **Q-Q Residuals** se observa una mejora con respecto a la normalidad de los errores, en **Scale-Location** se observa que hay homoscedasticidad, y en **Residuals vs Leverage** parece no haber outliers influyentes.

Table 3:

	Dependent variable:
	I(log(bpsystol))
bmi	0.009^{***} s.e: (0.002) t value: 6.161 Pr(>|t|): 2.41e-09
sex2	-0.049^{***} s.e: (0.015) t value: -3.311 Pr(>|t|): 0.00105
age	0.004^{***} s.e: (0.0004) t value: 8.403 Pr(>|t|): 1.95e-15
Constant	4.461^{***} s.e: (0.042) t value: 107.347 Pr(>|t|): < 2e-16
Observations	295
R^2	0.321
Adjusted R^2	0.314
Residual Std. Error	0.127 (df = 291)
F Statistic	45.922^{***} (df = 3; 291); p-value: < 2.2e-16
Note:	*p<0.1; **p<0.05; ***p<0.01



En el siguiente Cuadro, se muestra las pruebas de Shapiro-Wilk, Breusch-Pagan y Durbin-Watson, que plantean la hipótesis nulas de normalidad, homoscedasticidad y no autocorrelación, respectivamente. En todos los casos no hay evidencia suficiente para rechazar las hipótesis nulas.

	1
Normality (Shapiro-Wilk)	0.596
Homoscedasticity (Breusch-Pagan)	0.254
Autocorrelation of residuals (Durbin-Watson)	0.975

iii) Asociación entre masa corporal y presión arterial sistólica.

Se puede indicar que para una persona de cierta edad y sexo, tener un índice de masa corporal alto se asocia con una alta presión arterial sistólica. La relación es positiva o directa, es decir que a mayor índice de masa corporal, mayor es la presión arterial sistólica. En particular podemos notar que por cada unidad de incremento en la variable bmi, el log(bpsystol) aumenta en 0.009 para una persona con cierta edad y sexo. Esto es equivalente a decir que por cada unidad de incremento de mbi, bpsystol incrementa en $exp(0.009) = 1.009$ unidades.

La prueba de hipótesis nula es que $\beta_1 = 0$ y la alternativa que $\beta_1 \neq 0$, donde β_1 es el coeficiente o parámetro estimado asociado de la variable bmi. De acuerdo con la prueba $t - student$ tenemos un p-value asociado de $2.41e - 09$, lo cual es mucho menor a 0.05, con lo que rechazamos la hipótesis nula. Por lo tanto $\beta_1 \neq 0$ y es estadísticamente significativo al nivel de confianza del 95%.

iv) Gráfica resumen con la estimación puntual de la relación bpsystol y bmi.

A continuación presentaremos una gráfica resumen con la estimación puntual asociada a la relación entre bpsystol y bmi. Para esto consideremos sólo tres posibles edades: 30, 50 y 64, así como la diferenciación entre

mujeres y hombres. El comportamiento en general es que los hombres tienden a tener una mayor presión arterial sistólica, comparado con las mujeres. En todos los casos al aumentar la masa corporal, la presión arterial sistólica incrementa tanto para hombres como para mujeres. Además podemos observar que a mayor edad, es mayor la presión arterial sistólica tanto para hombres como para mujeres.