Regresión

Tarea 3

Christian Badillo

Tabla de contenidos

1 Conceptos.

1. ¿Qué son los residuales, cómo se construyen?

Los residuales se definen como el error entre el valor observado (y_i) y el valor predicho por el modelo ($\hat{y}_i$):

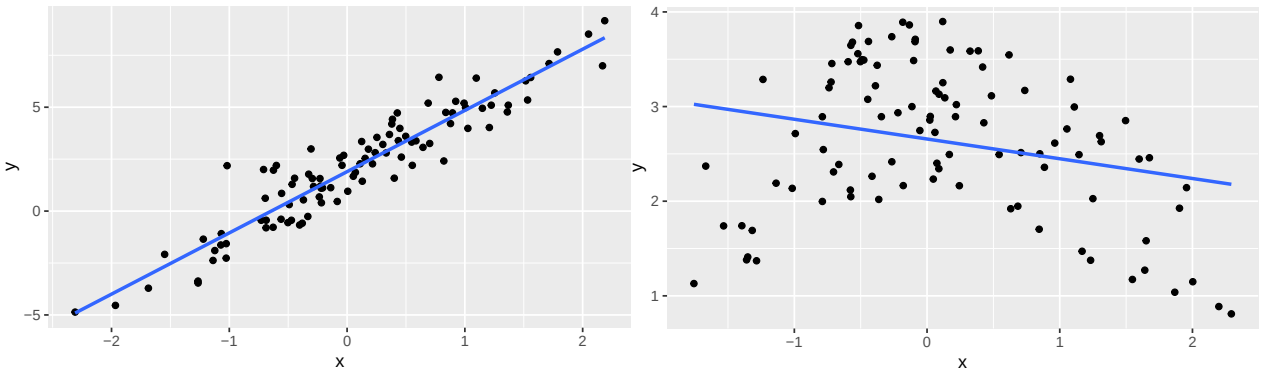
$$e_i = y_i - \hat{y}_i$$

2. ¿Cuál es la utilidad de los residuales respecto al modelo de regresión lineal?

Los residuales son útiles para evaluar la calidad del modelo de regresión lineal, ya que nos permiten identificar si el modelo es adecuado o no. Si los residuales son pequeños y no presentan patrones (aleatorios), entonces el modelo es adecuado, en otro caso puede que la relación entre las variables no sea lineal o que no se cumplan los supuestos del modelo.

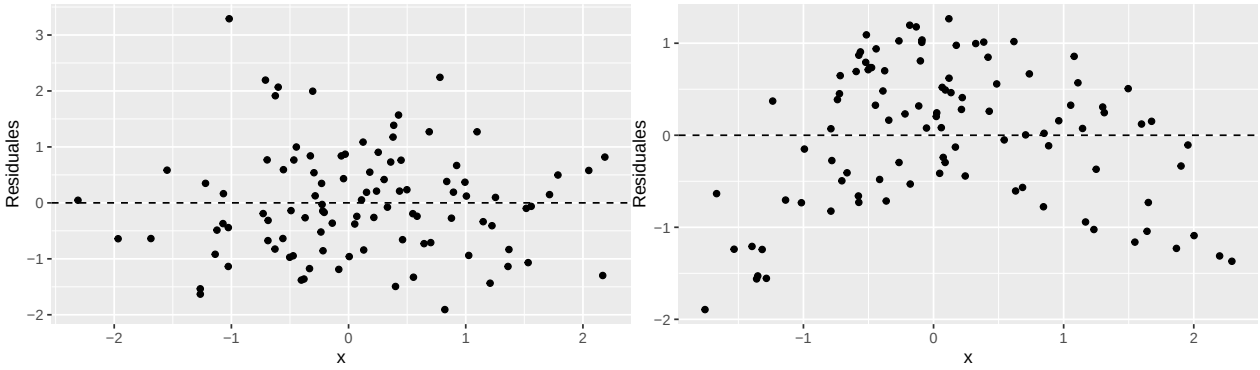
3. Da algunos ejemplos gráficos para representar las afirmaciones del inciso anterior.

Para esto se simularon dos conjuntos de datos, uno que cumple con los supuestos del modelo de regresión lineal y otro que no. En la figura 1 se muestra el gráfico de dispersión de los datos y en la figura 2 se muestra el gráfico de los residuales. Cómo se puede observar los residuales de los datos que cumplen con los supuestos del modelo son aleatorios y no presentan patrones, mientras que los residuales de los datos que no cumplen con los supuestos del modelo presentan un patrón muy claro, el cual indica que la relación entre las variables no es lineal.



(a) Datos que cumplen con los supuestos del modelo (b) Datos que no cumplen con los supuestos del modelo

Figura 1: Datos Simulados.



(a) Residuales datos lineales (b) Residuales datos no lineales

Figura 2: Gráficos de los residuales del modelo lineal.

2 Algebra Lineal.

1. Define el rango de una matriz, y da dos enunciados equivalentes, una con el rango y otra con su determinante. Explica a grandes rasgos por qué son equivalentes a que la matriz sea invertible.

El rango de una matriz es el número de columnas linealmente independientes que tiene. Tiene relación con el determinante de la matriz, ya que una matriz es invertible si y solo si su determinante es distinto de cero. Si el rango de una matriz es igual al número de columnas (o filas) de la matriz, entonces la matriz es invertible, ya que esto implica que las columnas (o filas) son linealmente independientes y por lo tanto el determinante es distinto de cero.

2. Describe(sin demostración) como interactúa el determinante con las operaciones de producto, transposición e inversión de matrices.

El determinante se relaciona con el producto de matrices mediante la siguiente propiedad:

$$\det(AB) = \det(A)\det(B)$$

En cuanto a la transposición de matrices, el determinante de una matriz es igual al determinante de su transpuesta:

$$\det(A) = \det(A^T)$$

Finalmente, en cuanto a la inversión de matrices, el determinante de una matriz invertida es igual al inverso del determinante de la matriz original:

$$\det(A^{-1}) = \frac{1}{\det(A)}$$

3. ¿Se puede afirmar algo respecto al determinante de la suma de matrices y la suma de los determinantes? ¿Hay alguna relación de igualdad o desigualdad que se satisfaga en general?

4. Define los siguientes conceptos, y da ejemplos de los mismos:

- Descomposición espectral de una matriz:

La descomposición espectral de una matriz es una factorización de una matriz cuadrada en términos de sus vectores y valores propios. La descomposición espectral de una matriz A se define como:

$$A = Q\Lambda Q^{-1}$$

- Matriz definida positiva:

Una matriz A es definida positiva si para todo vector no nulo x se cumple que $x^T Ax > 0$.

- Pseudoinversa de una matriz:

La pseudoinversa de una matriz A se define como la matriz A^+ que cumple con las siguientes propiedades:

$$\begin{aligned} AA^+A &= A \\ A^+AA^+ &= A^+ \\ (AA^+)^* &= AA^+ \\ (A^+A)^* &= A^+A \end{aligned}$$

- Descomposición polar de una matriz:

La descomposición polar de una matriz A se define como:

$$A = U\Sigma V^T$$

donde U y V son matrices ortogonales y Σ es una matriz diagonal con los valores singulares de A.

- Matriz ortogonal:

Una matriz A es ortogonal si cumple con la siguiente propiedad:

$$A^T A = I$$

3 Regresión Lineal.

Haz un análisis exploratorio de los datos , ¿qué variable tomarías como explicativa para intentar un modelo de regresión lineal?

Se exploró la relación entre las variables `density` y `residual sugar`, en la figura 3 se muestra la correlación entre estas dos variables. La correlación entre estas dos variables es de 0.36, lo cual indica que existe una relación lineal aunque débil entre estas dos variables. En la figura 4 se muestra el gráfico de dispersión de los datos, en el cual se puede observar que la relación entre las variables no es perfectamente lineal, dado que ciertos datos parecen seguir una relación cuadrática pero es probable que sean datos atípicos. Se decidio usar la variable `density` como variable explicativa para ajustar un modelo de regresión lineal, dado que es lógico pensar que a mayor densidad, mayor cantidad de azúcar residual.

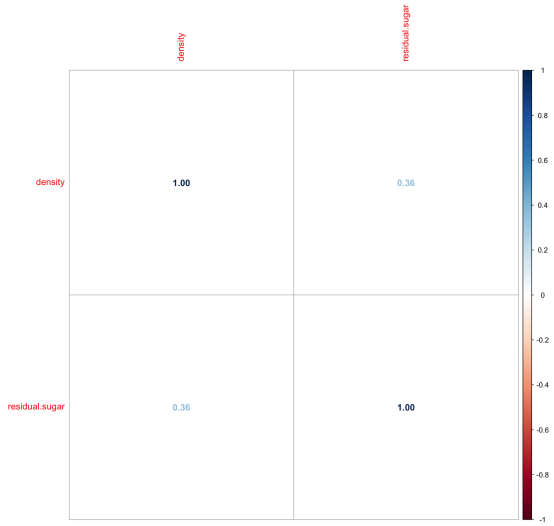


Figura 3: Matrix de correlación.

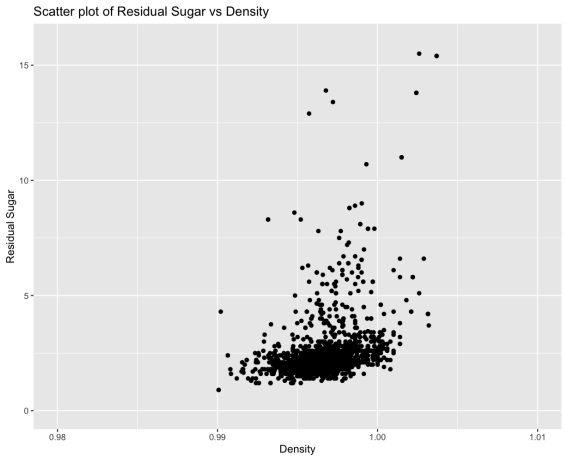


Figura 4: Gráfico de dispersión.

Tabla 1: Ajuste del modelo de regresión lineal.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-262.0113	17.41730	-15.04316	0
density	265.4135	17.47412	15.18895	0

2/3. Ajusta un modelo de regresión lineal de acuerdo a lo que hayas respondido en el inciso anterior. En el primer inciso seguramente notaste la presencia de outliers ¿Cómo afectan los mismos al modelo construido en el inciso anterior? Utiliza los residuales para apoyar tus afirmaciones.

Se ajustó un modelo de regresión lineal con la variable `density` como variable explicativa y `residual sugar` como variable respuesta. En la figura 5 se muestra el gráfico de los residuales del modelo de regresión lineal. En el gráfico de los residuales se puede observar que existen ciertos residuales que son muy grandes, lo cual indica que existen datos atípicos en el modelo, además de una patrón muy claro de los errores, lo cual da indicios que el modelo lineal no es un buen modelo para estos datos.

En la tabla 1 se resumen los coeficientes del modelo de regresión lineal, donde se observa que tanto el intercepto como el coeficiente de la densidad son significativos. El coeficiente de determinación del modelo es de 0.1262, lo cual indica que el modelo no explica mucho de la variabilidad de los datos y que no sería un buen modelo para predicciones. Igualmente se uso un gráfico cuantil cuantil para evaluar la normalidad de los residuales, en la figura 6 se muestra el gráfico, en el cual se puede observar que los residuales no siguen una distribución normal, lo cual es otro indicio de que el modelo no es adecuado. Además se analizó la autocorrelación de los residuales, pero no se encontró evidencia de autocorrelación en los residuales más allá del lag 3, por lo cual no se considera un problema en el modelo y podemos suponer que hay cierta independencia entre los residuales.

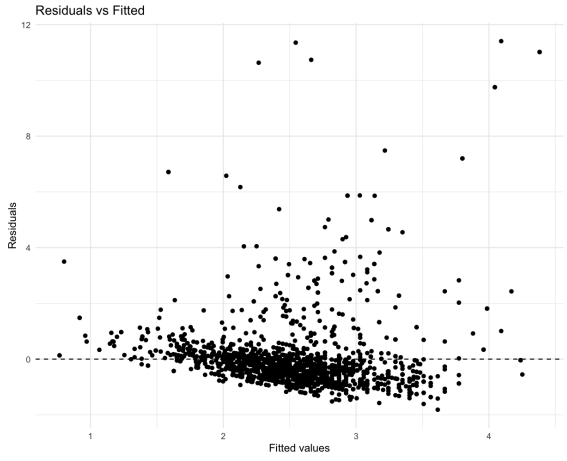


Figura 5: Residuales del modelo de regresión lineal.

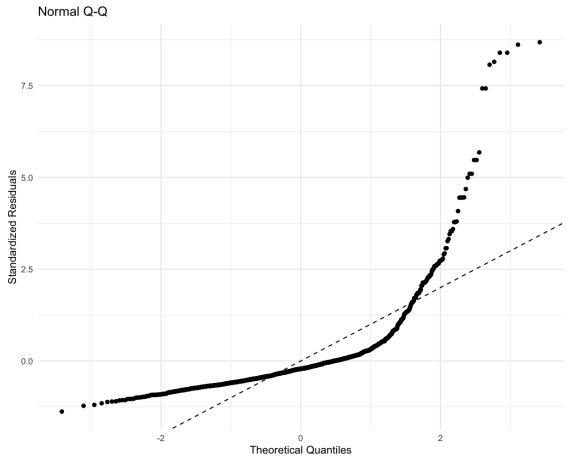


Figura 6: Gráfico QQ para los residuales del modelo de regresión.

4. Ajusta un nuevo modelo a los datos, pero eliminando los datos atíıpicos, ¿cómo cambia esto al modelo? ¿qué puedes concluir al respecto?

Se idéntificaron y eliminaron los valores atípicos usando el método del rango intercuartílico. En la figura 7 se muestra el gráfico de dispersión de los datos limpios, en el cual se puede observar que la relación entre las variables es más lineal que en los datos crudos. Se volvio a estimar la correlación en los datos, la cual aumento a 0.40 (véase figura 8).

Se ajustó un nuevo modelo de regresión lineal con los datos limpios, en la tabla 2 se muestran los coeficientes del modelo de regresión lineal donde se ve el cambio grande que hubo en la estimación del intercepto y la pendiente de la recta, por lo cual se puede concluir que los valores atípicos influyeron mucho en el modelo. El coeficiente de determinación del modelo de regresión lineal con los datos limpios es de 0.1564, lo cual indica que el modelo explica un poco más de la variabilidad de los datos, pero sigue siendo un modelo no adecuado para predicciones.

En la figura 9 se muestra el gráfico de los residuales del modelo de regresión lineal con los datos limpios, en el cual se puede observar que los residuales se comportan de una forma más aleatoria que cuando había datos atípicos pero se sigue notando un patrón en los residuales.

En el figura 10 se muestra el gráfico cuantil cuantil de los residuales donde se nota un mejor ajuste a una distribución normal que en el modelo anterior, pero aún se observan ciertos residuales que no siguen la distribución normal, especialmente en las colas de la distribución.

Tabla 2: Ajuste del modelo de regresión lineal con datos limpios.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-103.3287	6.506357	-15.88118	0
density	105.8665	6.528135	16.21696	0

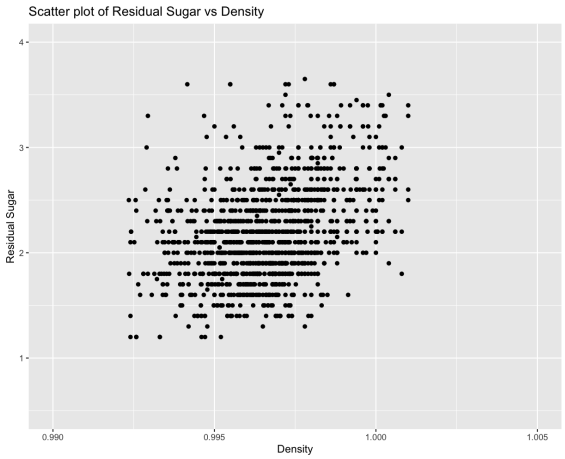


Figura 7: Gráfico de dispersión sin datos atípicos.

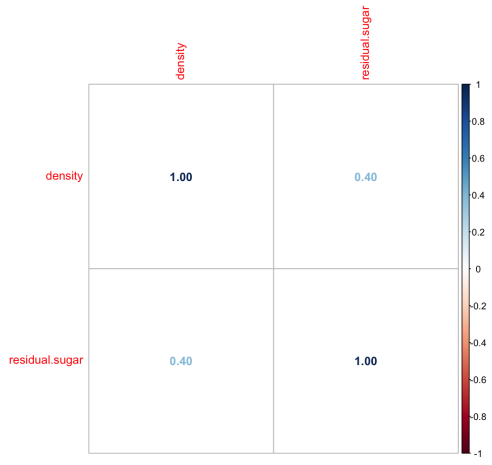


Figura 8: Matrix de Correlación.

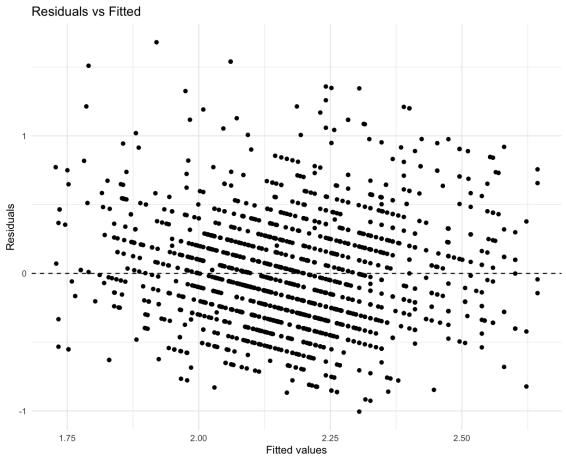


Figura 9: Residuales del modelo de regresión lineal con datos limpios.

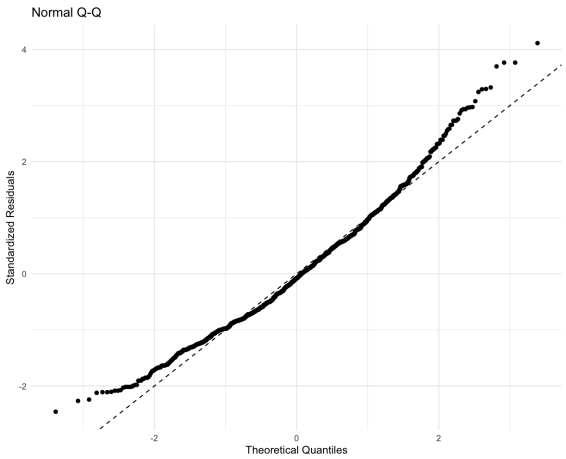


Figura 10: Gráfico QQ para los residuales con datos limpios.