ITESO, Universidad Jesuita de Guadalajara

Departamento de Matemáticas y Física

Ingeniería de características

Transformación de variables numéricas

Dr. Carlos Augusto Arellano Muro

Primavera 2022

Transformación: Es una función que mapea los elementos del conjunto X al conjunto Y.

- En términos estadísticos, se usan para estabilizar la varianza.
- Una vez trabajados los datos (usando una regresión, red neuronal, etc.), se realiza la transformación inversa para poder interpretar correctamente los datos.

1. Transformación logaritmo y recíproco

Función logaritmo: Una forma de visualizar la función logaritmo es como el inverso de la exponencial, es decir

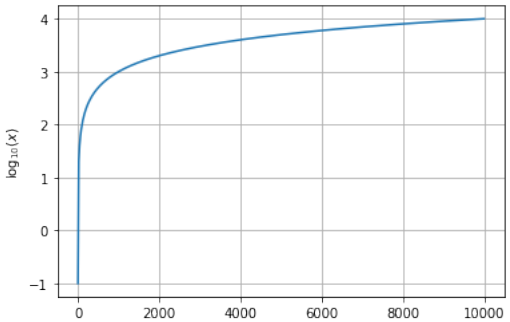
$$a^y = x \quad \leftrightarrow \quad \log_a(x) = y.$$

- Comprime el rango numérico alto y expande el rango bajo.
- Los valores pequeños, entre (0,1) los mapea al intervalo $(-\infty,0)$.
- Específicamente, la función $\log_{10}(x)$, mapea los intervalos mostrados en la Tabla 1 y el gráfico se muestra en la Figura 1.

Tabla 1: Mapeos de la función $\log_{10}$.

x	$\log_{10}(x)$
$[1, 10]$	$[0, 1]$
$[10, 100]$	$[1, 2]$
$\ldots$	

Figura 1: Gráfica de la función $\log_{10}(x)$.

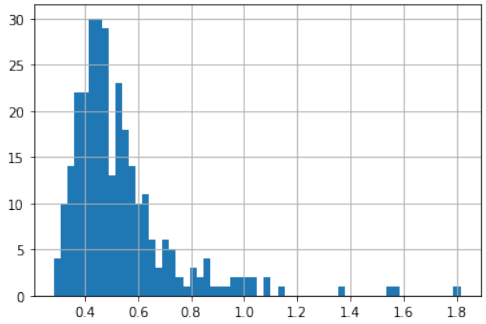


- Es útil para lidiar con números positivos con una distribución de cola pesada. La Tabla 2 muestra la sección inicial y final de datos numéricos ordenados con sesgo positivo, mientras que la Figura 2 muestra la densidad de distribución de estos datos.

Tabla 2: Fragmento de datos con cola pesada.

0	0.285
1	0.289
2	0.297
3	0.304
4	0.315
495	1.132
496	1.370
497	1.550
498	1.567
499	1.815

Figura 2: Histograma de los datos originales.

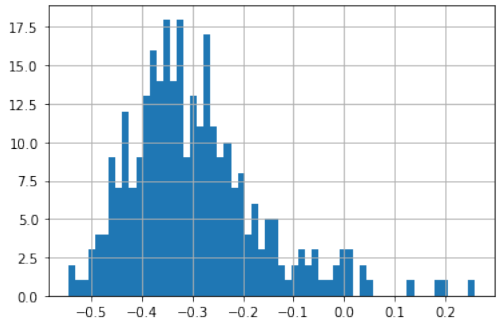


- Los datos originales se hacen pasar por la función logaritmo, resultando los mostrados en la Tabla 3, y su histograma en la Figura 3.

Tabla 3: Fragmento de los datos transformados.

0	-0.546
1	-0.539
2	-0.527
3	-0.517
4	-0.502
495	0.054
496	0.137
497	0.190
498	0.195
499	0.259

Figura 3: Histograma de los datos transformados.



Recíproco: Otra transformación útil para variables con datos con un sesgo muy grande es la función recíproca

$$y = 1/x$$

- Es más potente que la función logarítmica. Los valores grandes se atenúan más y los valores menores que 1 crecen más rápido.

- Tiene validez para número negativos. Si la variable presenta datos negativos, la función logaritmo arrojaría valores complejos, para evitar esto, una opción es usar la función reciproca, únicamente evitando el cero.

- A modo de comparación. Con valores menores que uno, por ejemplo $x = 0.1$, la función logaritmo regresa $\log(0.1) = -1$, que en magnitud es más pequeño que el obtenido con la función recíproco $10 = 1/0.1$. Además, al evaluar ambas funciones con $x = 10$, se obtiene $1 = \log(10)$ y $0.1 = 1/10$, donde se ve que los valores grandes son más atenuados con la función recíproco. La Figura 4 muestra los detalles entre el intervalo $[0.1, 10]$.

- Una sección de los datos transformados y la densidad de distribución se muestran respectivamente en la Tabla 4 y en la Figura 5.

Figura 4: Gráfica de la función recíproca.

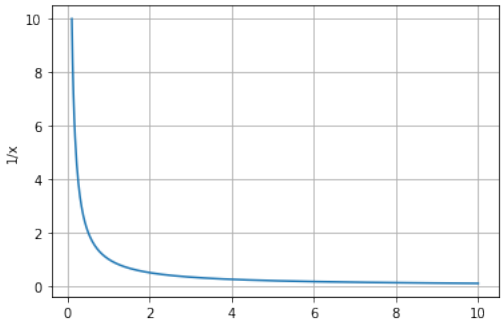
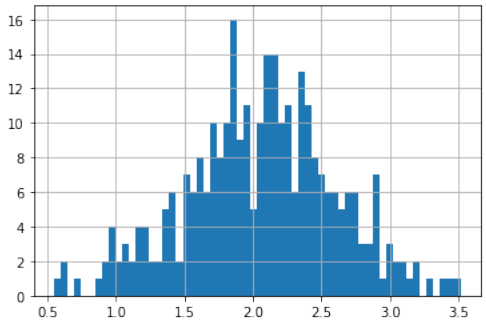


Tabla 4: Fragmento de los datos transformados por la función recíproca.

0	3.513
1	3.457
2	3.365
3	3.285
4	3.178
495	0.883
496	0.730
497	0.645
498	0.638
499	0.551

Figura 5: Histograma de los datos transformados.



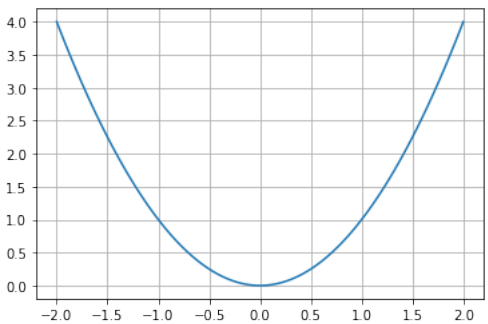
2. Transformación cuadrática y cúbica

Función cuadrática: Otra opción si se quiere compensar el sesgo positivo (cola a la derecha) es una transformación de la forma $y = \sqrt{x}$, ya que también amplifica los valores menores que uno y atenúa los mayores a éste. Ahora si se tienen datos cuya distribución presenta una cola cargada al lado izquierdo, se desea hacer lo contrario, y la forma inversa de la raíz cuadrada es la función cuadrática

$$y = x^2.$$

- Los valores pequeños, entre (-1,1) son atenuados, los valores mayores que uno, en magnitud, se amplifican. La Figura 6 muestra los detalles.

Figura 6: Gráfica de la función x^2.

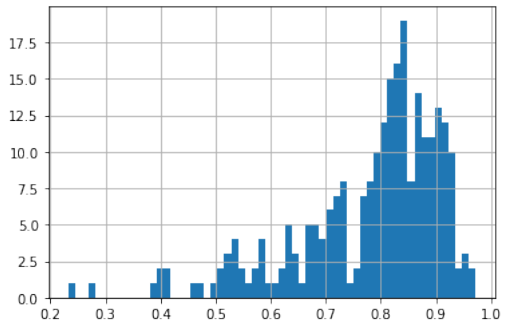


- Es útil para lidiar con valores que presentan un sesgo negativo (cola a la izquierda). La Tabla 5 muestra el inicio y el final de una distribución de datos cuyo histograma presenta un sesgo negativo, éste se muestra en la Figura 7.

Tabla 5: Fragmento de datos con sesgo negativo.

0	0.232
1	0.279
2	0.390
3	0.393
4	0.400
265	0.950
266	0.951
267	0.952
268	0.962
269	0.972

Figura 7: Histograma de los datos originales.

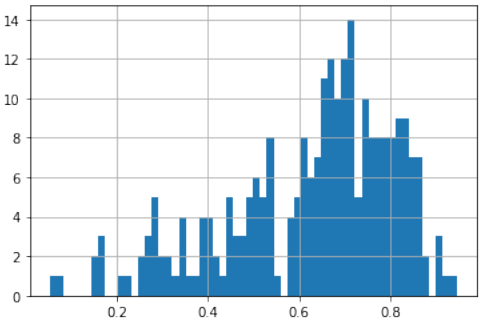


- Los datos originales se hacen pasar por la función cuadrática, resultando los mostrados en la Tabla 6, y su histograma en la Figura 8.

Tabla 6: Fragmento de los datos transformados por la función cuadrada.

0	0.054
1	0.078
2	0.152
3	0.154
4	0.160
265	0.903
266	0.904
267	0.907
268	0.926
269	0.945

Figura 8: Histograma de los datos transformados por x^2.



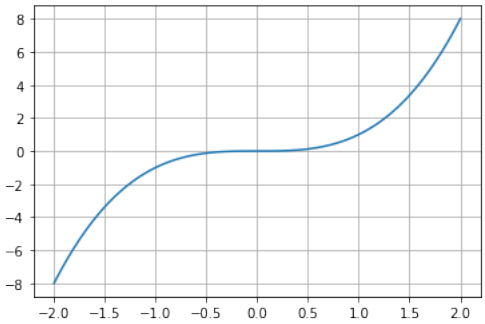
Función cúbica: Una transformación más de potencia. Tiene el mismo efecto que la función cuadrática: para variables con densidad de distribución que presentan una cola más alargada del lado izquierdo, las distribuye de forma que los valores pequeños sean más frecuentes. La transformación cúbica se expresa como:

$$y = x^3.$$

- Es más potente que la transformación cuadrática. Los números pequeños, al multiplicarse por ellos mismos, se hacen cada vez más pequeños, el resultante se atenúa aun más, es decir atenúa directamente el resultado de la función cuadrática. En cambio, para números mayores que uno, la función cuadrática se ve amplificada proporcionalmente por el valor a ser transformado.

- Mantiene el signo de los valores negativos. Se debe considerar que **los valores negativos, menores que menos uno, también crecerán en magnitud** por lo que se alejarán aún más hacia el lado izquierdo. La Figura 9 muestra el resultado de ser evaluada en el intervalo $[-2, 2]$.

Figura 9: Gráfica de la función x^3.

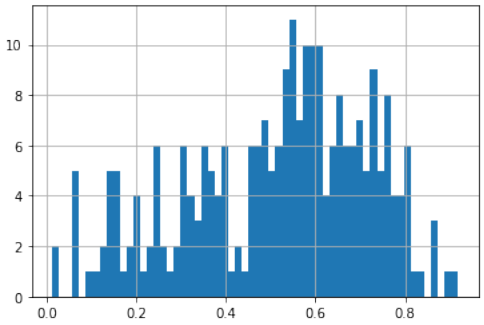


- Los primeros cinco y los últimos cinco elementos transformados se muestran en la Tabla 7, mientras que la representación gráfica de su frecuencia se muestra en la Figura 10.

Tabla 7: Fragmento de los datos transformados por la función cúbica.

0	0.013
1	0.022
2	0.059
3	0.061
4	0.064
265	0.859
266	0.859
267	0.863
268	0.890
269	0.918

Figura 10: Histograma de los datos transformados por x^3.



3. Transformación Box–Cox

Se habló de la transformación logaritmo y de las transformaciones de potencia, incluida la potencia de menos uno (recíproco) y la potencia de un medio (raíz cuadrada). Una generalización simple de estas transformaciones es la propuesta por los profesores David Cox y George Box

$$y = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases}$$

- Se define únicamente para valores de x positivos.
- Es una transformación flexible, si se escoge $\lambda = 0$ se tiene la transformación logarítmica.
- Con $0 < \lambda < 1$ se obtienen las transformaciones por raíces ($y = \sqrt{x}$, $y = \sqrt[3]{x}$, etc.).
- Al escoger $\lambda < 0$ se obtienen las transformaciones potencia de la función recíproca ($y = 1/x^2$, $y = 1/\sqrt{x}$, etc.), específicamente con $\lambda = -1$ se consigue la transformación recíproco.
- Finalmente, para $\lambda > 1$ la transformación se vuelve del tipo cuadrática o cúbica. Las gráficas con diferentes λ se muestran en la Figura 11.
- Como ejemplo, la Tabla 8 muestra los datos ordenados de una distribución cuya frecuencia se ve en la Figura 12.

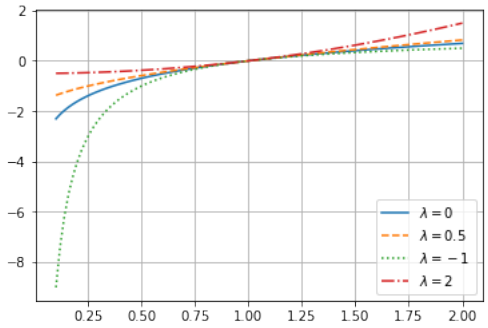
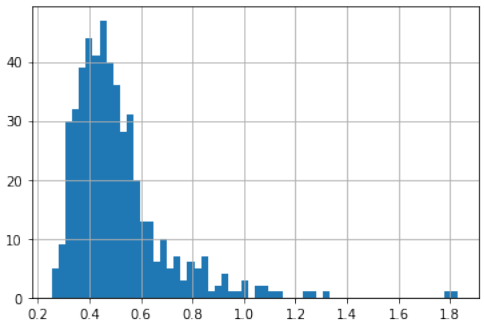


Figura 11: Gráfica de la transformación Box–Cox

Tabla 8: Fragmento de los datos sin transformar.

0	0.255
1	0.262
2	0.268
3	0.276
4	0.281
495	1.235
496	1.259
497	1.308
498	1.801
499	1.832

Figura 12: Histograma de los datos sin transformar.



- Intuitivamente se escoge $\lambda = 0$ para realizar una transformación logarítmica, ésta se muestra en la Figura 13.
- Los datos se transforman nuevamente con $\lambda = -0.9328$ obteniendo una distribución con una apariencia más cercana a la Gaussiana (Figura 14).

Figura 13: Histograma de los datos transformados por Box–Cox con $\lambda = 0$.

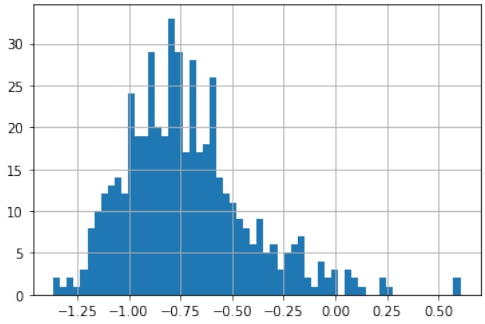
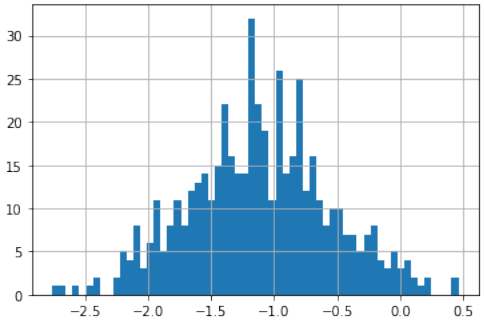


Figura 14: Histograma de los datos transformados por Box–Cox con $\lambda = -0.9328$.

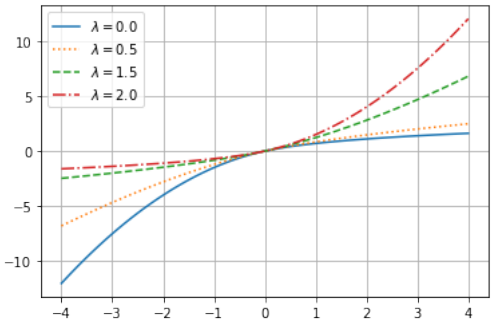


4. Transformación Yeo–Johnson

Una extensión de la transformación Box–Cox para variables con valores negativos es la transformación Yeo–Johnson [2]

- Siempre se puede desplazar la variable de forma que presente solo valores positivos para usar Box–Cox.

Figura 15: Representación gráfica de la transformación Yeo–Johnson.



- La transformación se define como

$$y = \begin{cases} \dfrac{(x+1)^{\lambda} - 1}{\lambda} & x \geq 0,\ \lambda \neq 0 \\ \ln(x+1) & x \geq 0,\ \lambda = 0 \\ -\dfrac{(-x+1)^{2-\lambda} - 1}{2-\lambda} & x < 0,\ \lambda \neq 2 \\ -\ln(-x+1) & x < 0,\ \lambda = 2 \end{cases}$$

- El primer criterio para identificar la función a ser ejecutada es el signo del elemento x después se usa λ bajo el mismo criterio que Box–Cox para estabilizar la varianza.
- Figura 15 muestra la evaluación de esta transformación en el intervalo $[-4, 4]$.
  - Note la continuidad entre valores negativos y positivos.
  - Tiene ambos efectos: Atenuar valores grandes ($\lambda = \{0.0, 0.5\}$) y amplificar valores grandes ($\lambda = \{1.5, 2.0\}$).
  - La función que se ejecuta cambia dependiendo del signo de x. Si se escoge $\lambda = 0$, la función con $x \geq 0$ será logaritmo, mientras que para $x < 0$ será cuadrática.
- Considere la densidad de distribución de la Figura 16, contiene valores negativos y un sesgo negativo, la Figura 17 muestra el efecto de la transformación con $\lambda = 3.074$

Figura 16: Densidad de distribución de una variable sin transformar.

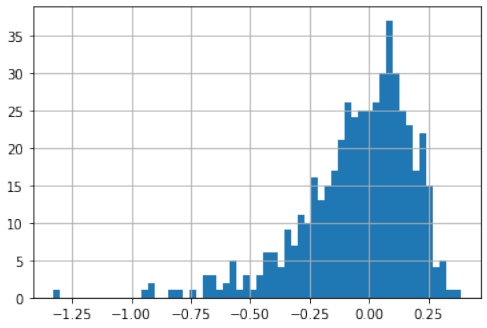
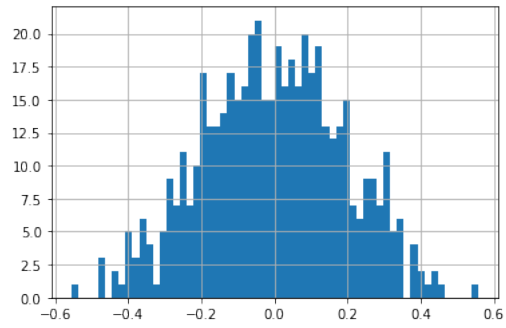


Figura 17: Histograma de la variable transformada con Yeo–Johnson usando $\lambda = 3.074$.



Referencias

[1] `https://www.cienciasinseso.com/transformacion-de-datos/`

[2] Yeo, I. K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. Biometrika, 87(4), 954–959.