Control 2

September 9, 2020

1 Introducción

En el presente documento se realizan un ajustes de regresión lineal a partir de datos de cierta ciudad obtenidos en un estudio que considera que el consumo de agua [m^2/mes] está relacionado con el consumo de energía eléctrica [kW/h].

Durante el proceso de análisis se descubre empíricamente que los datos originales no satisfacen del todo los supuestos del modelo de regresión lineal y se realiza una transformación para corregir dicha desviación.

El enfoque de este reporte se centra en la comparación de los ajustes lineales obtenidos para los datos originales y para los datos transformados, así como en la realización de inferencias a partir del último ajuste.

2 a) Grafique los datos y comente.

Primero, vamos a mencionar dos aspectos importantes de los datos con los que trabajaremos: los datos proporcionados constan de 50 pares de observaciones y son variables continuas.

Observación	Consumo eléctrico [kW/h]	Consumo de agua [m^2/mes]
1	0.16	3.76
2	0.44	4.05
3	0.63	4.26
4	0.76	4.37
5	1.04	4.51

A continuación se presenta una gráfica que presenta los pares de las mediciones `Consumo eléctrico` y `Consumo de agua`, medidos en *kw/hr* y m^2/mes respectivamente, de cada observación.

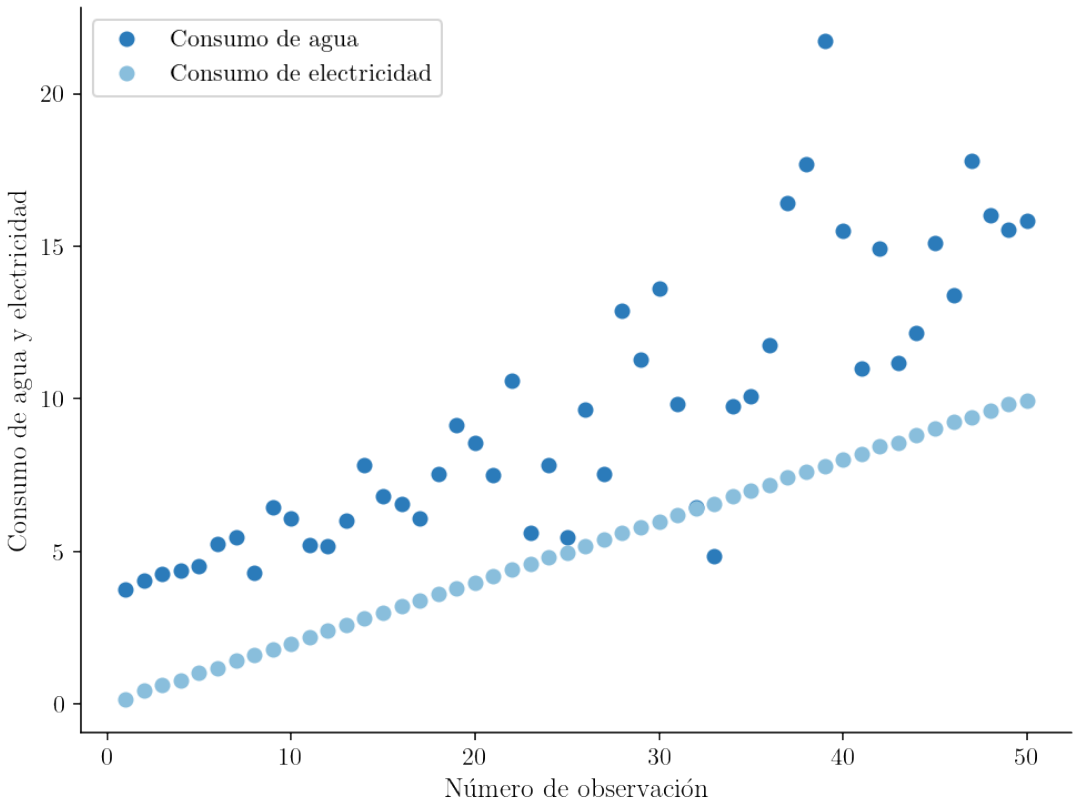

Figura 1. Consumo de agua medida en m^2/mes y consumo de electricidad medido en kw/hr

Como podemos observar, ambos tipos de consumo tienen una tendencia creciente. Sin embargo, el `consumo eléctrico` presenta cambios pequeños en cada observación, además, da la impresión de que el cambio es constante.

Por otro lado, a pesar de que el `consumo de agua` empieza con cambios pequeños y crecientes, después de la quinta observación podemos notar que se presentan cambios más grandes entre cada observación y no todos los consumos son mayores que los anteriores. En resumen, los valores del `consumo de agua` tienen más variabilidad que los del `consumo eléctrico`.

Ahora, vamos a graficar los datos tomando en cuenta al `Consumo eléctrico` como la variable independiente y al `Consumo de agua` como la variable dependiente.

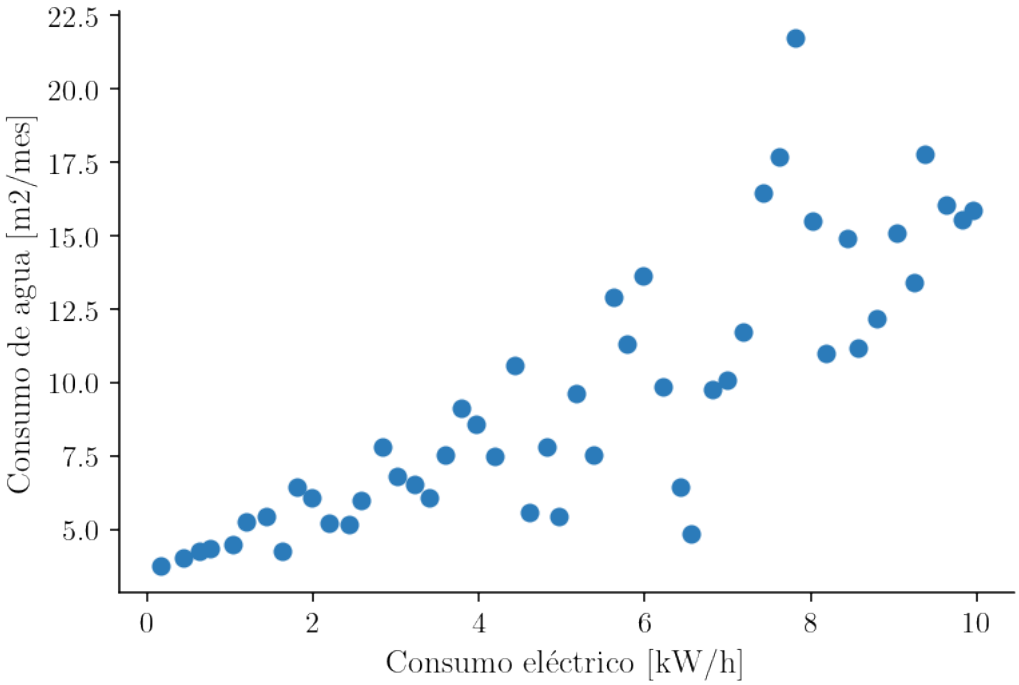

Figura 2. Gráfica de dispersión de las observaciones originales

3 b) Ajuste un modelo de regresión lineal simple sobre los datos sin transformar.

Ahora, queremos encontrar un modelo lineal que nos diga el `Consumo de agua` dado que tenemos un `Consumo eléctrico` fijo.

El modelo que obtenemos es:

y=1.3X+2.88

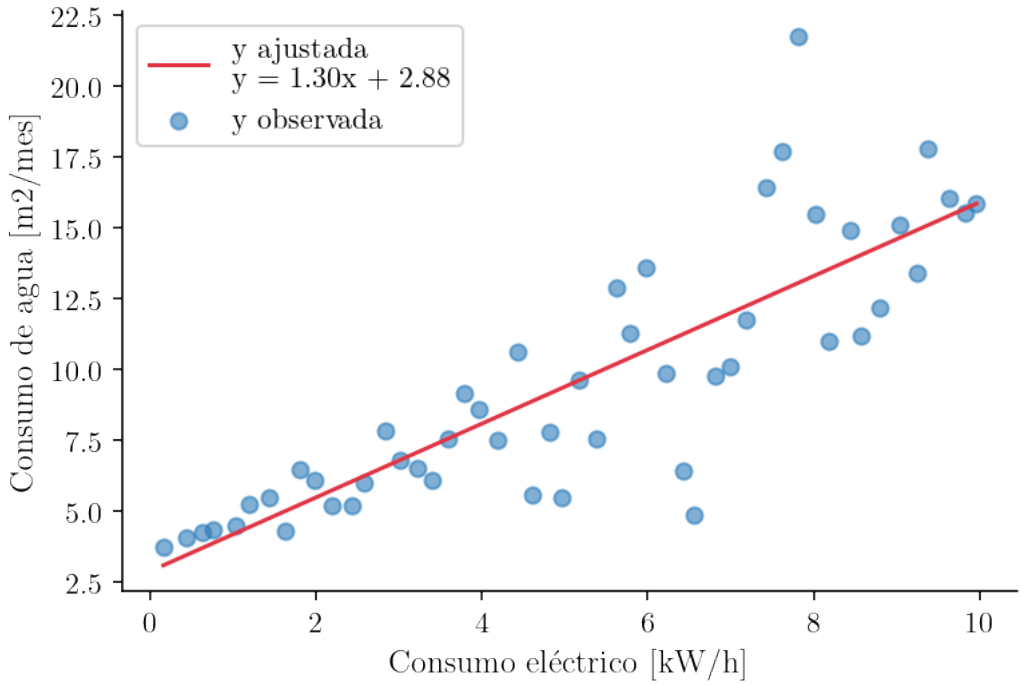

Figura 3. Gráfica de de dispersión de las observaciones originales con recta de ajuste de regresión lineal

```
Significancia del ajuste:
        Como el valor p = 1.68e-04 < 0.05 =  , se rechaza la hipótesis nula
H_0:  _0 = 0

Significancia del ajuste:
        Como el valor p = 1.88e-14 > 0.05 =  , se rechaza la hipótesis nula
H_0:  _1 = 0

El valor de R^2 para este ajuste es 0.7087
```

4 c) Verifique su modelo via análisis de residuales. Comente.

Con el modelo que obtuvimos, podemos calcular el consumo de agua esperado (respuesta ajustada) y el error residual entre el modelo y los datos observados (error residual). Ahora, vamos a graficar los `errores residuales` vs la `respuesta ajustada`.

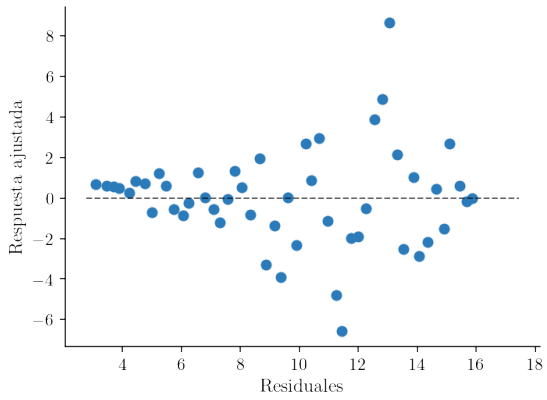

Figura 4. $\hat{e}$ vs. $\hat{y}$

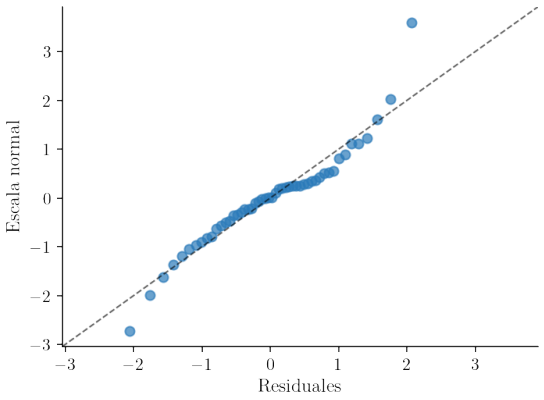

Figura 5. Grafica cuantil-cuantil

La figura 5 nos deja ver que nuestros residuales no cumplen con el supuesto de normalidad y además en la figura 4 podemos ver que los residuales tienen un comportamiento no aleatorio por lo que este modelo no tiene un muy buen ajuste.

5 d) Aplique la transformación de Box-Cox y construya un intervalo del 90% de confianza para λ. ¿Que valor de λ elegiría para la trasformación? Comente.

Aplicar la transformación de Box-Cox a la variable respuesta puede mejorar el ajuste y puede corregir violaciones de los supuestos del modelo de regresión lineal simple. Un requisito para poder aplicar esta transformación es que la variable respuesta conste únicamente de entradas positivas. Para validar que los datos disponibles satisfagan este requisito a continuación se cuenta la cantidad de entradas no positivas de la variable respuesta.

```
El número de entradas no positivas de la variable respuesta es: 0
```

Para poder aplicar directamente la fórmula de $y^{(\lambda)}$ de la transformación de Box-Cox, primero es necesario encontrar un valor óptimo $\hat{\lambda}$. El criterio de optimalidad para el parámetro $\hat{\lambda}$ es que la suma de los cuadrados de los residuales $SC_{\text{Res}}(\lambda)$ resultante de ajustar el modelo a partir de la respuesta sea mínima; es decir, λ está dado por

$$\begin{aligned}\hat{\lambda} &= \arg\min_{\lambda\in\mathbb{R}} SC_{\text{Res}}(\lambda) \\ &= \arg\min_{\lambda\in\mathbb{R}} \sum_{i=1}^{n} \left(y_i - y_i^{(\lambda)}\right)^2\end{aligned}$$

En la práctica, se elige un rango de valores de prueba del parámetro λ, se aplica la transfarmación de Box-Cox para cada uno de los valores dentro del rango, se ajusta el modelo para cada una de las transformaciones de la respuesta original y se utiliza aquél valor de λ con el que se obtenga la suma de cuadrados de los residuales $SC_{\text{Res}}(\lambda)$ mínima.

El resultado de aplicar este procedimiento con $\lambda \in [-1, 1]$ se muestra a continuación.

```
<pandas.io.formats.style.Styler at 0x128370510>
```

De los valores de λ dentro del rango seleccionado, se obtiene un valor mínimo de la suma de cuadrados de los residuales $SC_{\text{Res}}(\lambda)$ con *lambda* = -0.25. Es importante reslatar que la precisión de esta estimación depende de la cantidad de valores dentro del rango considerado, i.e. entre más fina sea la partición del intervalo de valores de prueba de λ, más precisa será la estimación de $\hat{\lambda}$.

Aunque probablemente haya un estimador más preciso de $\hat{\lambda}$, utilizar el valor *lambda* = -0.25 tiene la ventaja de la interpretación de que la transformación consiste en tomar el recíproco de la variable respuesta y aplicarle dos veces la función raíz cuadrada.

Para ver que λ = -0.25 es un valor adecuado para la transformación de Box-Cox, a continuación se obtiene gráficamente un intervalo de confianza para $\hat{\lambda}$. Para ello, se calcula el valor SC^* dado por

$$SC^* = SC_{\text{Res}}(\hat{\lambda}) \left(1 + \frac{t^2_{\left(1-\frac{\alpha}{2}, n-2\right)}}{n-2} \right)$$

y se utiliza para obtener el intervalo de confianza, que está dado por los valores de λ para los cuales $SC_{\text{Res}}(\lambda) \geq SC^*$.

```
<Figure size 432x288 with 0 Axes>
```

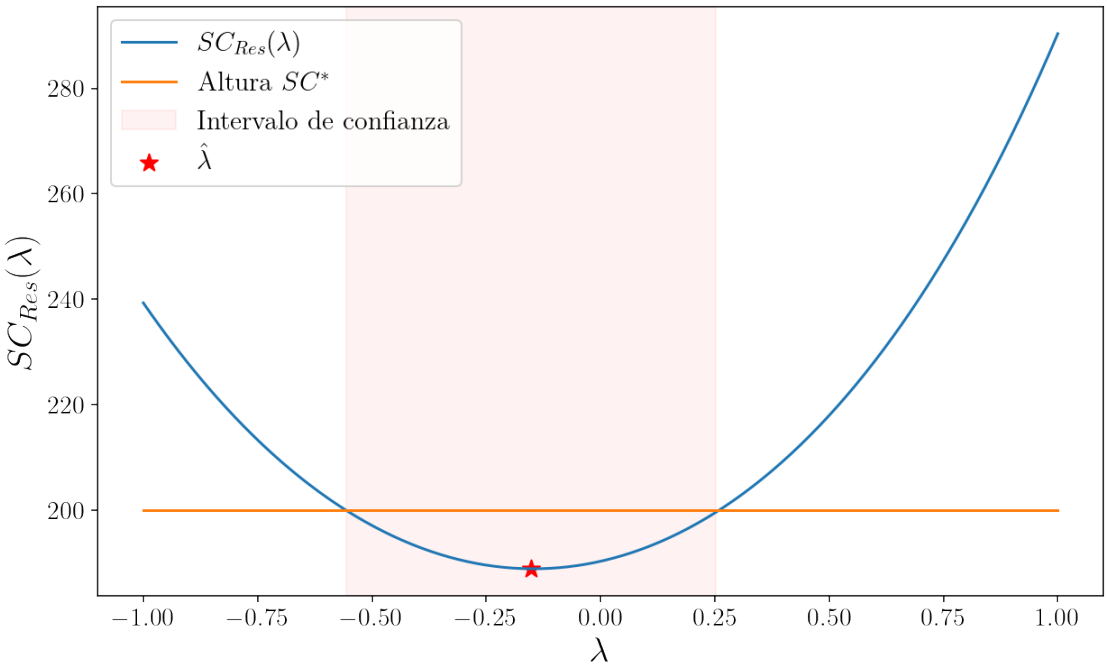

Figura 6. $SC_{Res}(\lambda)$ vs. λ con intervalo de 90.0

El punto importante a notar en la Figura X es que λ = -0.25 se encuentra dentro del intervalo de 90 % de confianza obtenido para $\hat{\lambda}$.

6 e) Grafique $y^{(\lambda)}$ vs. x. Comente.

A continuación se transforma la variable respuesta utilizando la transformación de Box-Cox con *lambda* = -0.25 y se grafican los resultados.

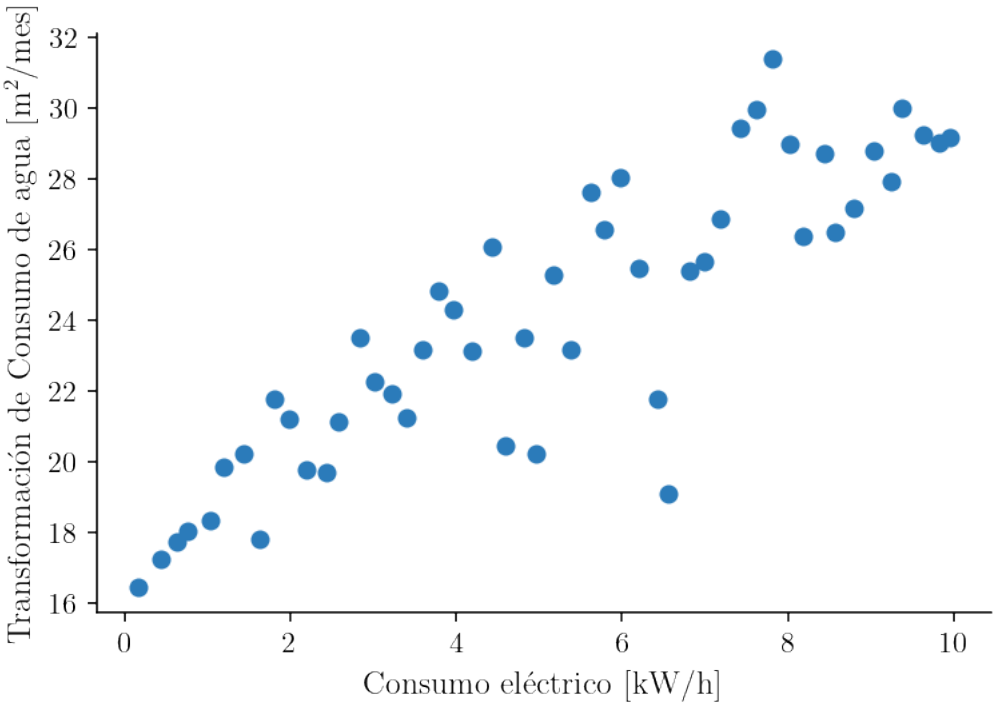

Figura 7. Gráfica de dispersión de las observaciones transformadas con λ = -0.25

En la Figura 7. es claro el desvanecimiento de la tendencia de aumento de la tasa de cambio de las observaciones de la variable respuesta con respecto a las observaciones del regresor. De hecho, los datos transformados demuestran una clara relación lineal.

7 f) Ajuste el modelo correspondiente y valídelo. Comente.

En esta sección se obtiene un ajuste de regresión lineal a partir de los datos cuya respuesta fue transformada por medio de la transformación de Box-Cox con *lambda* = -0.25. El ajuste obtenido, así como los parámetros de éste se encuentran graficados en la Figura 7.

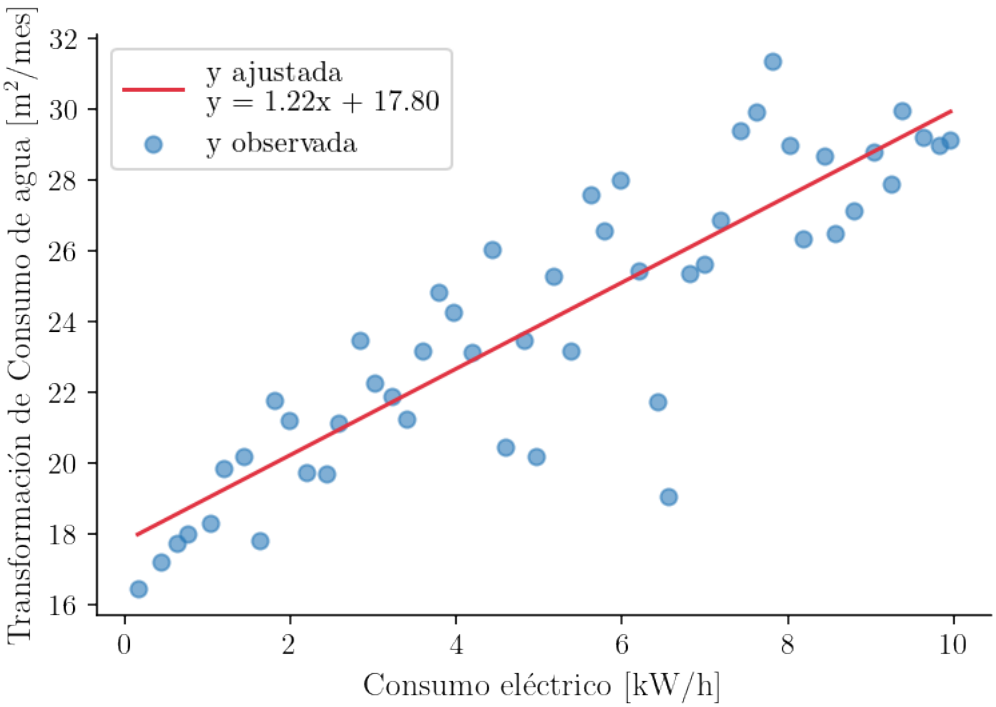

Figura 8. Gráfica de dispersión de las observaciones transformadas con λ = -0.25 con recta de ajuste de regresión lineal

A continuación se realizan las pruebas de hipótesis necesarias para validar la significancia del ajuste obtenido

```
Significancia del ajuste:
        Como el valor p = 0.00e+00 < 0.05 =  , se rechaza la hipótesis nula
H_0:  _0 = 0

Significancia del ajuste:
        Como el valor p = 1.11e-16 > 0.05 =  , se rechaza la hipótesis nula
H_0:  _1 = 0
```

Entonces, existe evidencia para afirmar que los datos transformados tienen un comportamiento linea.

También es posible evaluar la calidad del ajuste por medio del coeficiente de determinación R^2 que toma un rango de valores entre 0 y 1.

```
El valor de R^2 para este ajuste es 0.7653
```

Es importante resaltar que aunque el valor de R^2 obtenido para el ajuste de los datos trasformados sea mayor que el correspondiente a los datos originales, estas cantidades no son comparables en realidad debido a la transformación aplicada a la variable respuesta.

En la Figura 8. se encuentra la gráfica de los residuales determinados a partir del ajuste obtenido por medio de los datos transformados. Es destacable que la tendencia del error a aumentar conforme el incremento de la respuesta ajustada ha desaparecido y que para este caso los residuales aparentan tener una naturaleza más aleatoria que para el caso de los datos originales.

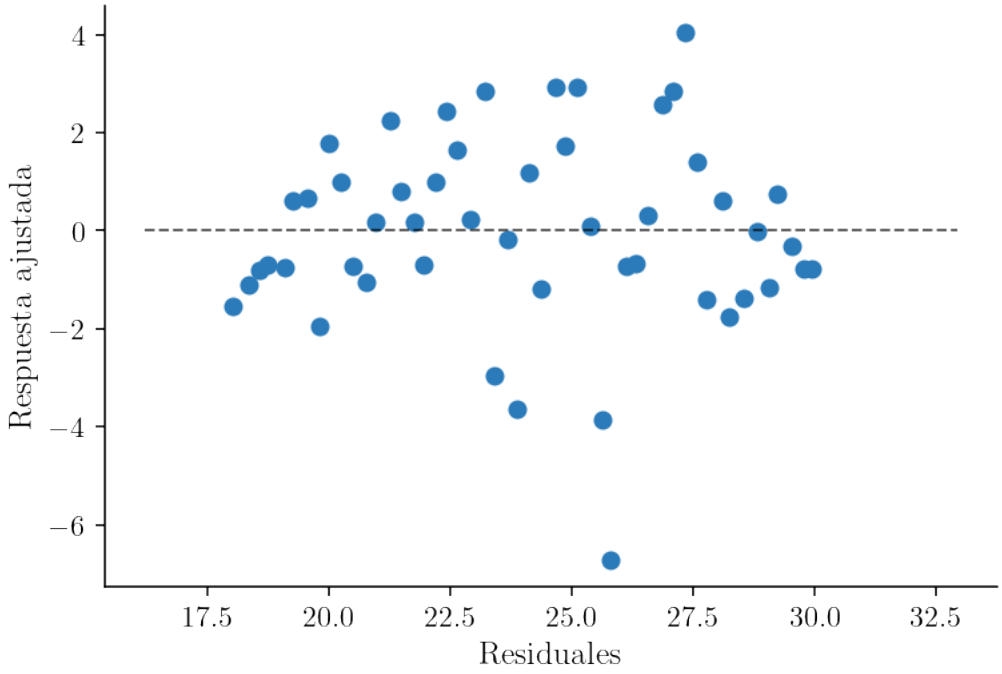

Figura 9. $\hat{e}$ vs. $\hat{y}$

De la Figura 9. vemos que con este nuevo modelo los residuales presentan un comportamiento más aleatorio que en el primero por lo que es un mejor modelo.

8 g) Construya un intervalo del 90% confianza para el consumo medio esperado si el consumo de energía eléctrica es de 7.57 kw/hr. (Nota: el intervalo es para el consumo de agua, no la respuesta transformada.).

Se consideran dos aproximaciones para la obtención del intervalo de confianza: 1. Obtenerlo a partir de los parámetros del ajuste obtenido con los datos originales 2. Obtenerlo a partir de los parámetros del ajuste obtenido con los datos transformados y aplicarle la transformación de Box-Cox inversa para obtener un intervalo en las unidades de los datos originales.

```
Intervalo obtenido a partir de los parámetros del ajuste original:
        El valor y_0   (11.98, 13.51) con 90.0 % de confianza para x_0 = 7.57

Intervalo obtenido a partir de los parámetros del ajuste transformado:
        El valor y_0   (11.08, 12.98) con 90.0 % de confianza para x_bar = 7.57
```

9 h) Construya un intervalo de predicción de 95 % para la demanda esperada si la generación de energía es de 5.1 kw/hr.

Nuevamente, se consideran las dos aproximaciones del inciso anterior.

```
Intervalo obtenido a partir de los parámetros del ajuste original:
        El valor y_0   (4.53, 14.52) con 95.0 % de confianza para x_0 = 5.1

Intervalo obtenido a partir de los parámetros del ajuste transformado:

El valor y_0   (5.33, 13.68) con 95.0 % de confianza para x_0 = 5.1
```

10 Conclusiones

A pesar de que el ajuste de regresión lineal obtenido a partir de los datos originales tuviera una apariencia adecuada, no era un ajuste válido debido a que los datos no satisfacían los supuestos del modelo.

Por medio de la transformación de Box-Cox pudimos obtener un modelo que se ajusta mejor a nuestros datos.