Informe de Estatística

Modelado Predictivo de la Temperatura Corporal para *Castor canadensis*

Fernández Costas, Xoán

Máster en Bioinformática para Ciencias de la Salud

Facultad de Informática, UDC

Curso 2023/2024

Índice:

1. Introducción. 1
   1.2. Ecología del Castor canadensis. 1
   1.3. Fisiología térmica. 2
   1.4. Introducción a los modelos de predicción. 2
   1.5. La regresión lineal múltiple. 3
2. Objetivos. 3
3. Metodología y resultados. 4
   3.1. Estudio descriptivo. 4
   3.2. Ajuste del modelo de regresión lineal múltiple. 7
      3.2.1. Diagnosis del modelo. 11
      3.2.2. Predicciones. 14
4. Discusión. 15
5. Conclusión 17
6. Bibliografía. 17

1. Introducción.

La predicción es una herramienta fundamental en una amplia gama de disciplinas, desde la meteorología hasta la epidemiología, pasando por las finanzas e incluso por la inteligencia artificial. En la actualidad, la predicción de variables ambientales y climáticas se ha vuelto esencial en numerosos campos de estudio. Este trabajo se centra en la creación y aplicación de un modelo de predicción dedicado a estimar la temperatura de un castor teniendo en cuenta la hora del día y si el castor tiene actividad fuera de su refugio o no. El conocimiento de la temperatura corporal de estos animales es crucial para comprender sus patrones de actividad, adaptaciones al entorno y comportamiento. Más adelante exploraremos la metodología utilizada para desarrollar este modelo de predicción. Este informe estadístico pretende mostrar cómo la predicción de la temperatura del castor puede proporcionar una visión más profunda de su comportamiento y fisiología, y cómo esto podría contribuir en un futuro al conocimiento y la preservación de esta especie emblemática. A continuación, se explicarán pequeñas pinceladas sobre la ecología y biología del castor, que permitirá una mejor comprensión de los objetivos de este trabajo. Del mismo modo se introducirán algunas pautas claves y conceptos que permitan entender el funcionamiento de un buen modelo predictivo.

1.2. Ecología del Castor canadensis.

El castor americano es uno de los roedores más grandes del mundo originario de América del Norte, que puede llegar a pesar hasta 30kg, poseyendo una cola plana y escamosa (Figura 1a). Los castores son animales nocturnos y edifican presas que sirven como refugio, proporcionando protección contra los depredadores y aislamiento contra el frío del invierno (Figura 1b) [1, 2]. La presa retiene el agua y hace cambiar el régimen de descarga de los arroyos. Esto da lugar a una serie de cambios ecológicos de vital importancia. La velocidad del agua disminuye y aumenta el área de suelo inundado, aumentando la retención de sedimentos y materia orgánica. Se reemplaza la fauna de invertebrados por una fauna de estanque, debido a la retención de agua, aumentando de esta forma 5 veces la biomasa de invertebrados [1, 2].

Por otra parte, la presencia de castores y la utilización de sus enormes dientes incisivos contribuyen a cortar la madera de los árboles de hoja caduca en invierno, causando la dominancia de las coníferas y modificando de esta manera el patrón de consumo de materia orgánica en el agua. A consecuencia de todo lo mencionado, los castores son denominados especies *keystones*, son especies que tienen un impacto desproporcionadamente grande en su ecosistema con relación a su abundancia, ya que alteran notablemente la estructura y dinámica de la corriente [2, 3].

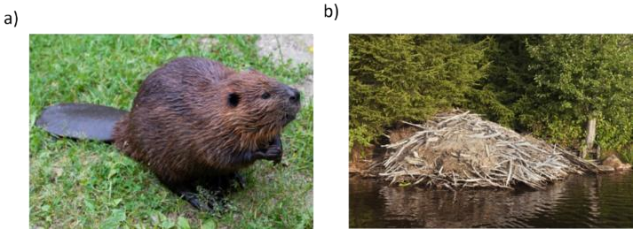

Figura 1: a) Imagen de una especie de Castor canadensis. 1b) Imagen de un refugio construido por castores.

1.3. Fisiología térmica.

La fisiología térmica de la mayoría de las aves y mamíferos se caracteriza por una considerable variación espacial y temporal de la temperatura corporal. La temperatura corporal es, por tanto, un parámetro clave en la investigación fisiológica, conductual y ecológica. La obtención de mediciones de temperatura en animales que se mueven libremente o que habitan en entornos naturales presenta desafíos significativos; no obstante, es factible llevar a cabo dicho proceso mediante la aplicación de diversas técnicas [4]. El muestreo de la temperatura interna puede realizarse mediante termometría, el empleo de registradores o transmisores implantados de forma quirúrgica, así como mediante dispositivos gastrointestinales o dispositivos no implantados mediante intervención quirúrgica.

La fisiología térmica de los vertebrados puede clasificarse ampliamente por su estabilidad en la temperatura corporal y la fuente de calor corporal. En nuestro contexto, al hablar de los castores americanos, nos centramos en la endotermia, que es el mantenimiento de una temperatura corporal homeotérmica alta y relativamente constante a través de una alta producción de calor metabólico que parece haber evolucionado por separado en linajes de aves y mamíferos [4]. La temperatura central de los mamíferos varía entre 30ºC en los monotremas y hasta 40º en otros grupos.

Común a todos estos grupos, la temperatura corporal es un parámetro fisiológico clave que proporciona información importante sobre el estudio de la termorregulación, la fisiología y el comportamiento o las respuestas al cambio ambiental [4, 5]. Es un error común suponer que existe un núcleo de referencia estándar o una temperatura corporal igual para cada especie y cada individuo [4, 5]. La temperatura de cualquier región anatómica es un producto de la producción de calor metabólico y el flujo sanguíneo dentro de esa región, incluida la tasa de calor perdido por los procesos físicos [5].

1.4. Introducción a los modelos de predicción.

Los modelos de predicción son herramientas o sistemas que utilizan datos históricos y relaciones matemáticas o estadísticas para estimar valores. Hay varios tipos de modelos de predicción, cada uno diseñado para aplicaciones específicas. Para elucidar el objetivo de este trabajo es preciso conocer uno de los modelos de predicción más populares, como son los modelos de regresión. El análisis de regresión es una técnica estadística para determinar la relación entre una sola variable dependiente (respuesta) y una o más variables independientes (explicativas). Estos modelos buscan establecer relaciones entre variables para predecir un valor específico. De tal forma que, en este modelo se asume que la variable dependiente es el efecto de la variable independiente, de tal forma que el valor de los predictores se utiliza para estimar o predecir el valor más probable de la variable respuesta.

1.5. La regresión lineal múltiple.

La regresión lineal múltiple es una técnica de análisis estadístico utilizada para predecir una variable dependiente a partir de dos o más variables independientes. A diferencia de la regresión lineal simple, que se enfoca en una sola variable independiente, la regresión lineal múltiple considera múltiples factores para predecir con mayor precisión el valor de la variable objetivo.

En la ecuación del modelo lineal general de regresión, la variable dependiente se representa como una combinación lineal de las variables independientes, y la relación se expresa de la siguiente manera:

$$Y = m(X_1, \cdots, X_k) + \varepsilon = m(\boldsymbol{X}) + \varepsilon$$

Donde:

- Y es la variable respuesta
- **X** es el vector de variables explicativas
- ε es el error aleatorio

De esta manera suponemos que la variable respuesta (Y) y las explicativas están relacionadas linealmente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_p X_p + \varepsilon = \boldsymbol{X}'\beta + \varepsilon$$

Donde ahora β representa los coeficientes de regresión asociados a las variables predictoras.

El objetivo de la regresión lineal múltiple es determinar los valores de los coeficientes que minimizan la suma de los cuadrados de los residuos, es decir, que produzcan la mejor estimación de la variable dependiente. Esto se logra mediante técnicas estadísticas como el método de mínimos cuadrados.

2. Objetivos.

En este trabajo se propuso construir un modelo que permita predecir la temperatura corporal del castor americano teniendo en cuenta la hora del día y si el castor tiene actividad fuera de su refugio. En concreto se persiguieron dos objetivos específicos:

- Realizar un modelo de predicción en R para predecir la temperatura corporal del castor americano.
- Comprobar si este modelo cumple las hipótesis estructurales necesarias para que el modelo funcione correctamente.

3. Metodología y resultados.

La metodología de este informe estadístico se centrará en primer lugar en el estudio descriptivo de los datos extraídos de la base de datos de R, concretamente se trata de un *data.frame*, denominado *beaver data* [6, 7]. Esta base de datos forma parte de un estudio desarrollado en Wisconsin sobre la regulación de la temperatura corporal de los castores (*Castor canadensis*) y realizado por Reynolds en 1994 [6]. La información se obtuvo de cuatro castoras adultas atrapadas *in vivo*, a las que se sometieron a una cirugía para implantarles un transmisor de radio sensible a la temperatura. Tras la cirugía, las lecturas fueron tomadas cada 10 minutos. Al mismo tiempo, se registró la ubicación de los castores y su nivel de actividad se dividió en dos categorías dependiendo si el castor se encontraba dentro del refugio o fuera de él, suponiendo que las actividades de alta intensidad sólo ocurren fuera del refugio. La base de datos cuenta con dos subconjuntos *beaver1* y *beaver2*, para la realización de este informe estadístico sólo se tendrán en cuenta los datos del subconjunto *beaver2*.

Todo el conjunto de datos presenta 100 filas y 4 columnas, en las que se registró:

- *Day*: Día de la observación. Los datos incluyen únicamente el día 307 e inicios del día 308, que corresponde a los días 3 y 4 de noviembre de 1994. Variable cuantitativa discreta [6].
- *Time*: La hora del día formateada como hora-minuto. Variable cuantitativa continua [6].
- *Temp*: La temperatura corporal en grados Celsius. Variable cuantitativa continua [6].
- *Activ*: Indicador de actividad. 1 indica que el castor se encuentra fuera del refugio y, por tanto, se encuentra realizando actividades de alta intensidad. Variable cuantitativa discreta [6].

Descrito ya toda la información subyacente al conjunto de datos con el que se va a trabajar en este informe, comenzaremos a realizar todo el estudio estadístico con RStudio, un entorno de desarrollo integrado (IDE) ampliamente utilizado para programar en R. Cabe mencionar que la hora del día formateada como hora-minuto, al cambiar el día se reinicia, es por ello por lo que para este estudio estadístico se reformateó la hora del día para que se tratase de una variable cuantitativa continua sin saltos temporales y no dé lugar a fallos en los resultados. Así mismo, tras pasar el día 307, las horas venideras se contarán desde las 24h en adelante desde la realización del estudio.

3.1. Estudio descriptivo.

El estudio descriptivo se centra en la recopilación y presentación de datos para describir un fenómeno. Su objetivo es proporcionar una imagen detallada y objetiva de los datos que se están estudiando. Mediante diversas técnicas estadísticas y visualizaciones, el estudio descriptivo permite ofrecer una comprensión clara de las características, patrones y tendencias presentes en los datos, lo que es esencial para la toma de decisiones informadas.

En un contexto geográfico y temporal, los datos fueron obtenidos en los días 3 y 4 de noviembre en Wisconsin donde el ocaso se produce sobre las 17h y 45 minutos aproximadamente. En primer lugar, en los datos obtenidos (*beaver2*) existe un claro aumento de la temperatura cuando el castor se encuentra realizando actividades de alta intensidad (*activ* = 1), que se correlaciona con sus hábitos nocturnos (Figura 2). Según los datos del *data.frame* sobre las 16h las castoras empiezan a realizar actividades de alta intensidad hasta las 2 de la madrugada. No obstante, desde las 9h y 30 minutos hasta las 16h los animales se encuentran en reposo (*activ* = 0) en su madriguera.

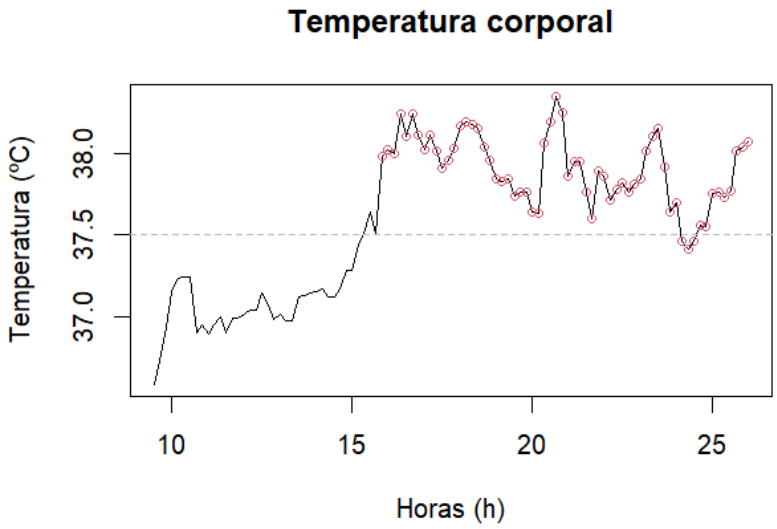

Figura 2: Representación de la temperatura corporal frente al tiempo, comenzando desde las 9:30 h del día 307 hasta las 2:00 h del día 308. Los datos resaltados mediante un círculo en color rojo corresponden a la información recopilada de los castores que están participando en actividades físicas intensas.

A la vista de los resultados se crearon dos subconjuntos a partir del *data.frame* con el fin de analizar cada una de las condiciones (activos o en reposo) por separado. De lo contrario, la estimación de estadísticos como la media o la mediana se verá desviada por los diferentes valores de cada una de las condiciones. En la recogida de datos se obtuvo 38 muestras de las castoras en reposo y 62 de los castores realizando actividades de alta intensidad (Figura 3a). En cuanto a la temperatura del castor en reposo se sitúa en torno a los 37.10 ºC, mientras que su mediana se encuentra en 37.09 ºC. Comprobamos como los valores se agrupan en torno a los 37 ºC y siempre manteniéndose en un rango estable gracias a la termorregulación y su capacidad endotérmica (Figura 3c). Haciendo referencia a la Figura 3b se aprecia una drástica subida de la temperatura hacia el final de las 15h, probablemente debido a que los animales estén empezando a prepararse para salir del refugio (recordemos que el código binario de la actividad física del animal se rige por la presencia de este dentro o fuera de la madriguera).

Por el contrario, si nos fijamos en la temperatura media de las castoras fuera del refugio, en general existe un aumento de la temperatura en los datos obtenidos. En este caso la temperatura media ha aumentado a 37.90ºC y su mediana a 37.91ºC (Figura 3c). Este hecho radica en un incremento de la actividad física de estos animales, lo que da lugar a una mayor actividad muscular, generando calor como subproducto del metabolismo

llevado a cabo por el ejercicio. Del mismo modo, aumenta el flujo sanguíneo permitiendo distribuir el calor generado en los músculos a través del cuerpo.

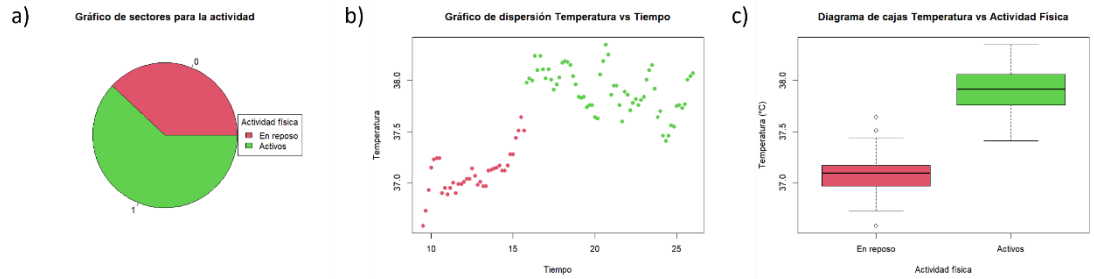

Figura 3: Representación gráfica de los datos: 3a) Gráfico de sectores para la actividad física, comprobando que hay más datos de castores activos que en reposo. 3b) Gráfico de dispersión de la temperatura corporal de los castores frente al tiempo, donde los puntos rojos señalan los datos de los castores en reposo, mientras que los verdes los datos de los castores activos. 3c) Diagrama de cajas de la temperatura corporal de los castores en reposo y cuando se encuentran ejerciendo actividad física, se muestra como la temperatura media aumenta considerablemente. Se observan 3 datos atípicos en el estado de reposo.

Nos fijamos que, a priori, la temperatura de los castores cuando están activos tiene una mayor varianza que cuando se encuentran en reposo. Así mismo, se calculó la desviación típica (DT) para cada uno de los subconjuntos, en el caso de los datos en reposo la DT es de 0.20ºC, mientras que la DT en el caso de los datos activos es de 0.21ºC. Para evaluar y visualizar la distribución de nuestros datos se realizó conjuntamente un histograma y un gráfico de densidad (Figura 4), que nos permite visualizar la distribución de los datos, donde podemos apreciar dos grupos e identificar dos modas (picos en el gráfico), correspondientes a las diferentes temperaturas acorde a las condiciones de la actividad física en el tiempo.

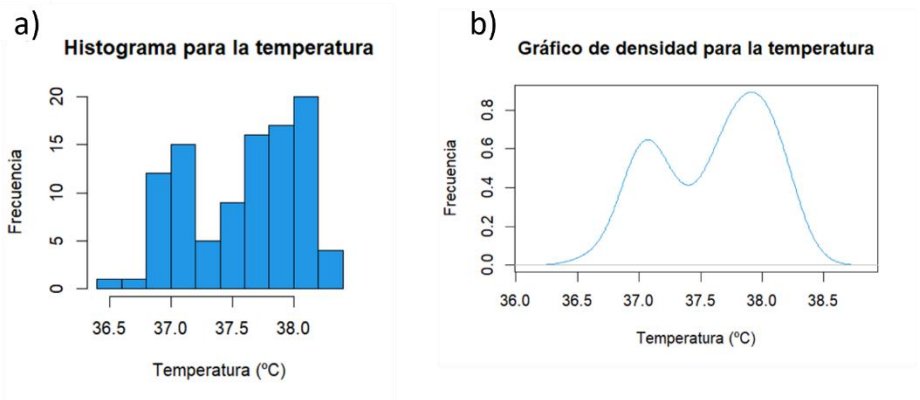

Figura 4: Representación gráfica de los datos: 4a) Histograma para la temperatura y 4b) Gráfico de densidad para la temperatura. Se puede comprobar en ambos gráficos, que los puntos de mayor frecuencia coinciden con las medias de temperatura para los diferentes estados de actividad física.

Pese a que las gráficas muestren un resultado negativo respecto a la distribución normal de los datos. A mayores se realizó el test de normalidad de Shapiro-Wilk, esta es una prueba utilizada para evaluar si una muestra proviene de una población con una distribución normal. Utilizando la función *shapiro.test()* en R se obtuvo un p-valor = $7.764 \cdot 10^{-5}$, considerando un nivel de significación del 10%, da como resultado el rechazo de la hipótesis nula y se asume que claramente hay suficientes evidencias significativas de que la variable temperatura (*temp*) no sigue una distribución normal.

Para finalizar con el estudio descriptivo, se propuso estudiar cada una de las relaciones entre las diferentes variables. Con este objetivo se diseñó una matriz de diagramas de dispersión con Rstudio (Figura 5), donde se pudo observar que las variables temperatura y actividad física tienen una fuerte tendencia lineal positiva. Cabe destacar que la gráfica de dispersión de los datos de la actividad física se ubica solamente entre los valores de 0 y 1, por su característica binaria.

A pesar de ello, podemos observar como en el pico más bajo de la temperatura del castor coincide cuando la actividad física es 0 y el pico más alto cuando la actividad física es 1, exponiendo cierta correlación entre los valores.

a)

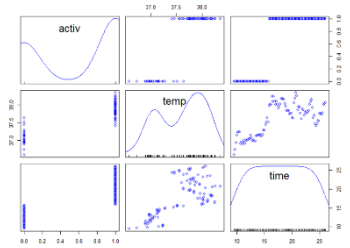

Figura 5: Representación de la correlación de los datos. 5a) Matriz de gráficos de dispersión, se comparan las variables de actividad física, tiempo y temperatura corporal entre sí.

3.2. Ajuste del modelo de regresión lineal múltiple.

Ajustar un modelo de regresión lineal múltiple implica encontrar los coeficientes óptimos para las variables predictoras con el fin de predecir la variable respuesta. Recordamos que nuestro objetivo es predecir la temperatura de los castores, teniendo en cuenta la hora del día y el hecho de si el castor tiene actividad fuera o dentro de su refugio. En primer lugar, para realizar el ajuste del modelo se estimarán los parámetros de este mediante el método de mínimos cuadrados, estos estimadores mínimos cuadráticos minimizan la suma de los cuadrados de las diferencias entre los valores reales y las predicciones de la respuesta. De esta manera la solución es el estimador mínimo cuadrático de β:

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{Y}$$

Estos estimadores coinciden con los estimadores máximo-verosímiles, ya que tratan de buscar el valor del estimador más cercano posible al valor verdadero. Este ajuste se realizó a través de R, con la función *lm()* donde se obtuvo el siguiente modelo de regresión lineal ajustado:

$$Temperatura = 37.286 - 0.015 \cdot Tiempo + 0.931 \cdot Actividad + error$$

Al ver el resumen del modelo con la función de *summary()* (Tabla 1), podemos comprobar el coeficiente de determinación (R^2) ajustado que tiene un valor de 0.77, sugiriendo a priori que puede ser un buen modelo de predicción para hallar la temperatura de los castores. Aproximadamente el 77.8% de la variabilidad en la variable dependiente es explicada por el modelo de regresión múltiple. El p-valor es muy significativo ($< \alpha = 0.01$) en el caso del intercepto y el estimador de la variable *activ*, no obstante, para un nivel de significación del 5% el p-valor de β_2 no es significativo.

	Estimación	Pr(>|t|)
β_0	37.286	< 2e-16 ***
β_1	0.015	0.0662 .
β_2	0.931	< 2e-16 ***

R^2 Ajustado
0.778

Tabla 1: Se muestran los valores de la estimación y el p-valor correspondientes a cada uno de los estimadores en el contexto del modelo de regresión múltiple. Asimismo, se presenta el R^2 ajustado del modelo, con un valor de 0.778. Hay que destacar que, los estimadores β_0 y β_2 tienen un p-valor muy pequeño, en contraste con el estimador β_1.

En este punto, se llevaron a cabo contrastes de hipótesis individuales sobre los parámetros. Específicamente, se realizó un contraste individual de la t para contrastar que, ninguno de los parámetros pueda llegar a ser igual a cero. Se calcularon los límites de los intervalos de confianza para un $\alpha = 0.05$ (Tabla2).

	2.5%	97.5%
β_0	37.073	37.499
β_1	-0.031	0.001
β_2	0.798	1.064

Tabla 2: Se representan los valores de los límites de los intervalos de confianza para un nivel de significación del 95%. Se observa que el intervalo de confianza de β_1 contiene al 0.

Los resultados mostrados en la Tabla 2, indican que los intervalos de confianza no son efectivos para un nivel de significación del 5%, debido a que el intervalo de confianza del estimador β_1 contiene al 0. Esto sugiere que no hay evidencias significativas para rechazar la hipótesis nula de que el valor verdadero del parámetro es igual a 0. Por lo que la variable *time* podría no estar contribuyendo de manera significativa al modelo.

Se propuso ampliar el intervalo de confianza para un nivel de significación del 10% (Tabla 3). Teniendo en cuenta que, los nuevos intervalos para los estimadores serán menos precisos, pero tendrán un menor riesgo de no contener el verdadero valor del parámetro.

	5%	95%
β_0	37.107	37.463
β_1	-0.028	-0.002
β_2	0.798	1.064

Tabla 3: Se representan los valores de los límites de los intervalos de confianza para un nivel de significación del 90%.

Al realizar un nuevo intervalo menos restrictivo, se proporciona un rango plausible de valores para los verdaderos parámetros. Con este cambio, todas las variables son estadísticamente significativas distintas de 0, aceptando de este modo la hipótesis alternativa del contraste individual de la t. En nuestro contexto, sugiere que las variables asociadas a los distintos estimadores tienen un impacto significativo en la variación de la temperatura corporal de los castores.

Al graficar los resultados podemos ver una cierta tendencia lineal de nuestro modelo entre los valores predichos y observados de la variable respuesta (Figura 6a).

Los residuos dentro de este contexto son las diferencias entre los valores observados y los valores predichos por el modelo. En otras palabras, un residuo es la discrepancia entre la respuesta real y la respuesta predicha por el modelo de regresión. Los residuos son esenciales para evaluar la calidad del ajuste del modelo y para verificar si se cumplen las suposiciones del modelo de regresión lineal, como la normalidad de los residuos, la homocedasticidad, y la independencia de los residuos, que se analizarán más adelante. Se graficaron los valores predichos frente a los residuos (Figura 6b). El gráfico presenta un patrón aleatorio de los datos y una dispersión de los residuos continua, indicando a priori linealidad y homocedasticidad en el modelo de regresión lineal múltiple. Cabe destacar que en la Figura 6 se representan dos conjuntos de datos agrupados, pertenecientes a cada uno de los dos valores que se presentan en la variable *activ*.

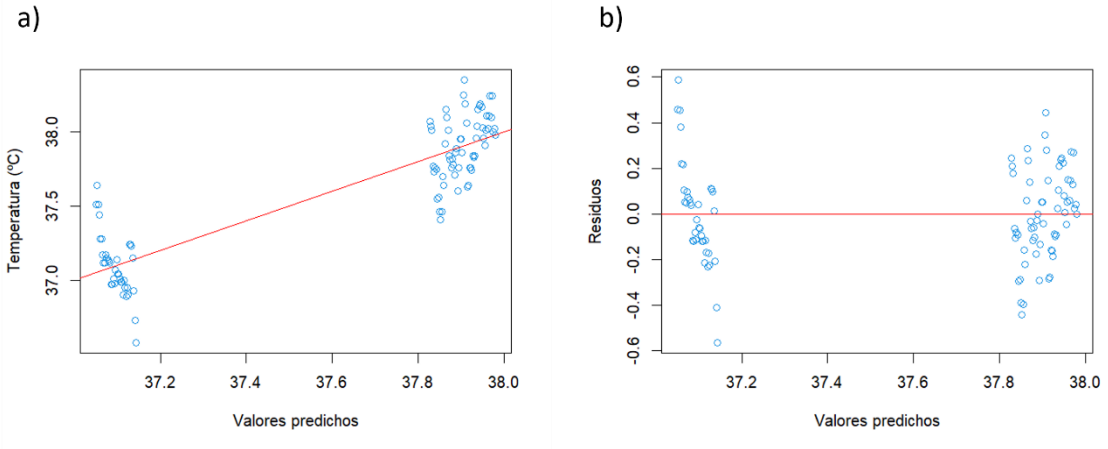

Figura 6: 6a) Gráfico de dispersión de los valores predichos frente a los observados de la variable respuesta. 6b) Gráfico de dispersión de los valores predichos frente a los residuos. Se asume homocedasticidad y linealidad.

Es importante contrastar que no existe relación lineal entre la variable respuesta y las variables explicativas. En otras palabras, si los coeficientes de regresión son significativamente diferentes de cero, indica la existencia de una relación significativa entre las variables predictoras y la variable respuesta. Este contraste se denomina contraste de regresión de la F y se resuelve por el análisis de la varianza en regresión lineal múltiple. Al realizar este contraste en R obtenemos para el tiempo un p -valor = 0.066 y para la actividad física un p-valor $< 2.2 \cdot 10^{-16}$. Por lo que hay evidencias claramente significativas de que existe un efecto lineal en las variables explicativas, suponiendo un nivel de significación del 10%. Se rechaza por tanto que el modelo lineal no es significativo.

Continuando con el estudio del modelo de regresión lineal múltiple en R, calculamos los residuos estandarizados, que son los residuos, divididos por la desviación estándar de los errores del modelo. Al estandarizar estos residuos, se obtiene una medida adimensional que nos facilita la comparación entre los conjuntos de datos y son fundamentales para el diagnóstico de problemas como la multicolinealidad, la heterocedasticidad y la falta de linealidad. Su fórmula es la siguiente:

$$r_i = \frac{e_i}{\hat{s}_R\sqrt{1 - h_{ii}}} \sim N\,(0,1)$$

En la Figura 7, se representan múltiples gráficos para los residuos estandarizados. Tras la normalización de los residuos podemos comprobar como las figuras pertinentes, siguen una distribución normal. Así mismo, el diagrama de cajas nos ofrece una visualización de la distribución y la dispersión de los residuos estandarizados (Figura 7c). Al interpretar estos gráficos nos damos cuenta de que se observan datos atípicos que podrían influir en el modelo y se analizarán más adelante. A mayores observamos una simetría en la distribución de los residuos (Figura 7a).

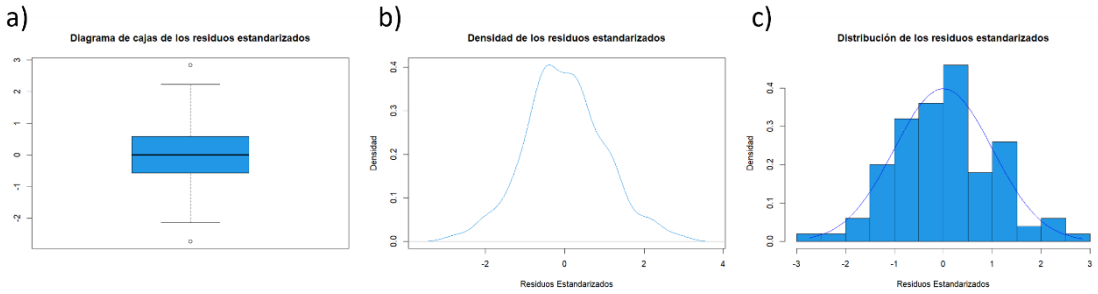

Figura 7: Representación de los residuos estandarizados. 7a) Diagrama de cajas, se comprueba que la caja es simétrica y se destacan dos datos atípicos. 7b) Densidad de los residuos estandarizados, se aprecia que los residuos estandarizados siguen una distribución normal. 7c) Distribución de los residuos estandarizados, en esta ocasión se representa un diagrama de cajas y la línea azul representa una función gaussiana que se asemeja a la distribución que siguen los residuos estandarizados.

A continuación, en la Figura 8a se representa el gráfico de dispersión de los residuos estandarizados frente al valor de influencia o *laverage*. Este tipo de gráficos permiten identificar las observaciones influyentes en nuestro modelo estadístico. Concretamente el "*Leverage*" de una observación mide cuánto se aleja el valor de la variable predictora con esa observación del valor medio de las variables predictoras. Debido a esto, las observaciones con valores extremos en las variables predictoras tienen un mayor valor de influencia. Visualizando el gráfico podemos ver diferentes puntos enumerados que se corresponden a aquellos residuos grandes y con alto *leverage* que afectan significativamente a nuestro modelo. En concreto sería necesario investigar más a fondo los datos 37, 38 y 1 para comprender su influencia y decidir si deben ser excluidos. En esta ocasión, no haría falta excluirlos debido a que el modelo de regresión lineal múltiple es bastante bueno y se demostrará más adelante no se verá afectado por estos tres valores. No obstante, al analizar estos valores, concretamente el 37 y 38, se destaca la presencia de una alta temperatura cuando supuestamente el castor está en reposo, a pesar de esto recordemos que la actividad física se rige por la presencia del castor en la madriguera. Es por ello, por lo que el castor puede estar en movimiento dentro del refugio reflejando de esta manera los valores atípicos. En la Figura 8b se presenta el gráfico escala-localización, donde se evalúa la homocedasticidad de los residuos y se comprueba nuevamente la presencia de los 3 valores atípicos vistos anteriormente, el 1, el 37 y el 38.

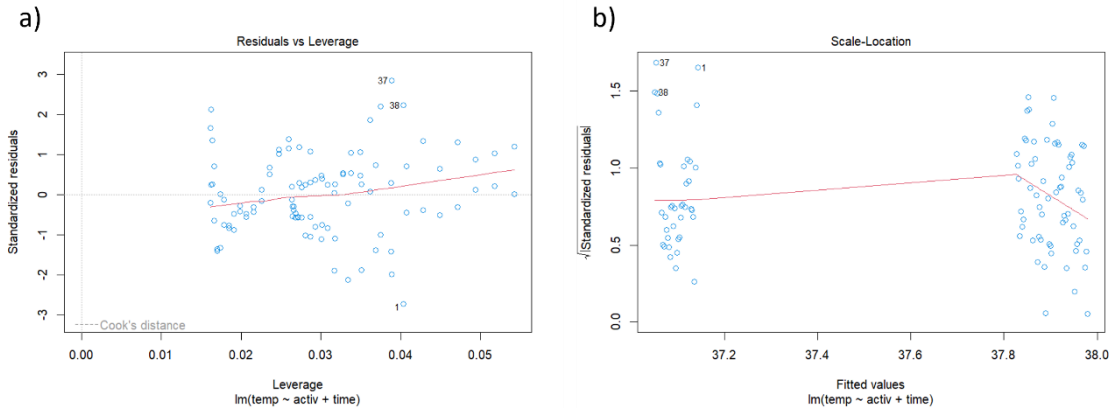

Figura 8: Se representan dos gráficos. 8a) Gráfico de los residuos frente al Leverage. 8b) Gráfico de escala-localización. Los dos gráficos en conjunto muestran un patrón aleatorio de los datos y una dispersión de los residuos continua, indicando homocedasticidad en los mismos. Asimismo, en las dos representaciones gráficas se evidencian tres puntos de datos que ejercen una notable influencia, identificados como los datos 1, 38 y 37.

Posteriormente, precedemos a graficar un Q-Q plot (*Quantile-Quantile plot*), útil para evaluar si los datos siguen una distribución específica y para identificar desviaciones de la normalidad (Figura 9). En un Q-Q plot se comparan los cuantiles de la muestra de datos con los cuantiles esperados bajo la distribución teórica asumida (en nuestro caso la distribución normal), como vemos en la siguiente figura los puntos se distribuyen aproximadamente a lo largo de una línea diagonal, representando que los datos siguen una distribución normal. A mayores, también se realizó la prueba de normalidad de Shapiro-Wilk, dando lugar a un p-valor de 0.833, que, asumiendo un valor de $\alpha = 0.10$, no rechazamos la hipótesis nula debido a que no hay suficientes evidencias estadísticamente significativas de que nuestros residuos estandarizados no sigan una distribución normal.

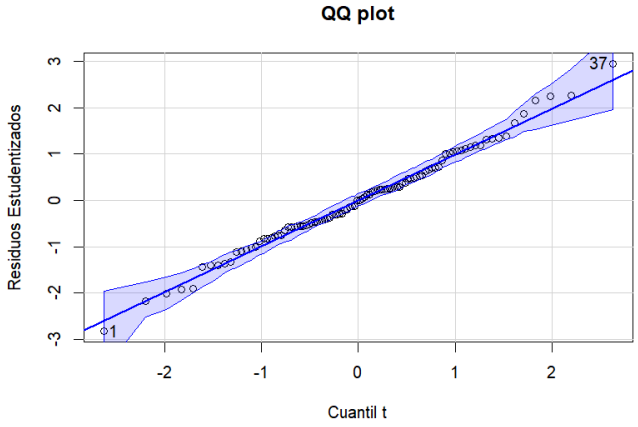

Figura 9: Representación de un gráfico que ilustra una adecuada concordancia entre los valores observados y los valores esperados según el modelo.

3.2.1. Diagnosis del modelo.

Llegados a este punto, se dispuso a realizar una serie de pruebas cualitativas para determinar y confirmar los diferentes atributos de los residuos estandarizados. Hay que comprobar que las hipótesis básicas del modelo de regresión lineal múltiple se cumplen. En caso de que alguna de estas hipótesis no sea cierta, las conclusiones obtenidas no son fiables y pueden ser erróneas. Estas hipótesis básicas del modelo de regresión múltiple son: linealidad, normalidad, homocedasticidad, independencia y comprobar si se tiene un problema de alta multicolinealidad.

Cabe destacar, que se ha de tener siempre en cuenta que el nivel de significación utilizado para los diferentes contrastes es del 10% en concordancia con el nivel de significación fijado con anterioridad en los límites de los intervalos de confianza.

En primer lugar, se realizó el test de Lilliefors, este test es una prueba de bondad de ajuste que permite señalar si los datos analizados siguen una distribución específica. En R con la función de *lillie.test()* se determinó un p-valor de 0.683, lo que da lugar a la aceptación de la hipótesis nula, e indica que los datos analizados parecen seguir una distribución específica, esta como hemos comprobado con anterioridad es la distribución normal. Posteriormente se realizó la prueba de normalidad de Cramer-von Mises. Esta prueba permite evaluar la bondad de ajuste de una muestra a una distribución teórica continua, en nuestro caso sigue siendo una distribución normal. Con la función *cvm.test()* en R se obtuvo un p-valor de 0.468, asumiendo nuevamente que los datos siguen una distribución normal.

Por otro lado, se realizó el contraste de homocedasticidad de Breusch-Pagan, que permite evaluar la varianza en un modelo de regresión y cuya hipótesis nula refleja que la varianza de los errores del modelo es constante a lo largo de todos los niveles de las variables explicativas. Con el código *bptest()* en R, se calculó un p-valor = 0.118, indicando claramente que no hay evidencias estadísticamente significativas de que los datos analizados no sean homocedásticos. Tras esto se realizó un contraste de residuos atípicos con ayuda del test de Bonferroni, que, con la utilización de una distribución t de student indica el estado atípico de los residuos estudentizados. Por lo que la presencia de un valor significativo indica la presencia de un valor atípico extremo. Con el cálculo en R y la función *outlierTest()* se determinó un p-valor ajustado (p de Bonferroni) para el dato número 37 de 0.391, demostrando que no existe ningún valor atípico en los datos que pueda influir de manera significativa a la temperatura corporal de los castores.

Para visualizar la relación entre la variable dependiente y las variables independientes se realizó un gráfico de variables añadidas o adicionales (Figura 10). En este caso se muestra la temperatura corporal de los castores frente al tiempo y a la actividad física. Como vemos se puede identificar un cierto patrón lineal entre las dos variables y la temperatura, dando lugar a una clara interacción entre la variable de interés y las variables independientes. Con esta representación podemos comprobar cómo el modelo lineal múltiple es el adecuado para la predicción de la temperatura corporal de los castores a través del tiempo y la actividad física.

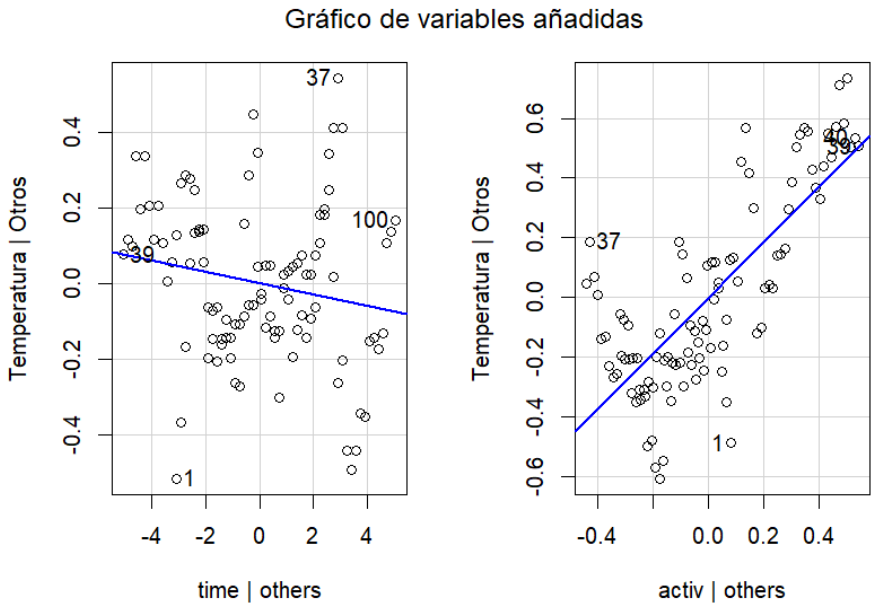

Figura 10: Gráfico de variables añadidas. En cada uno de los gráficos se representa la temperatura frente al tiempo (Gráfico de la derecha), y frente a la actividad física (Gráfico de la izquierda).

Con el fin de contrastar la aleatoriedad de nuestros datos, se realizaron dos contrastes de aleatoriedad. El primero es la prueba de Durbin-Watson, con la función de R *durbinWatsonTest()* que calculó un p-valor = 0, teniendo en cuenta un valor de $\alpha = 0.10$, se rechaza la hipótesis nula, mostrando que existen evidencias estadísticamente significativas de que hay autocorrelación serial en los residuos. Lo mismo ocurre para la segunda prueba de aleatoriedad que es el test de Box-Ljung con la función en R *Box.test()* se calculó un p-valor $< 2.2 \cdot 10^{-16}$. Corroborando nuevamente, al igual que la primera prueba, que los residuos consecutivos no son independientes y que hay indicios de autocorrelación.

También se realizó una prueba estadística no paramétrica para evaluar una vez más la aleatoriedad en una secuencia de datos, en donde la hipótesis nula de este contraste denominado la prueba de Runs, indica si los datos siguen un patrón aleatorio. Con la función de *runs.test()* en R, se generó un p-valor $= 4.61 \cdot 10^{-10} < \alpha$ (0.10). Con este resultado la hipótesis nula fue rechazada y se aceptó que los datos no siguen un patrón aleatorio. No podemos asumir que los datos sean aleatorios.

A mayores, se buscó verificar la homocedasticidad de los datos para comprobar si la varianza de los residuos de nuestro modelo de regresión es constante. Por ello se realizó un contraste en R con la función *ncvTest()* dando lugar a un p-valor $= 0.24 > \alpha$ (0.1), demostrando que no hay evidencias suficientemente significativas para demostrar que la varianza de los residuos no es constante.

En la figura 11 se representa un gráfico de dispersión de los residuos estandarizados absolutos frente a los valores ajustados, en él se incluye una línea rosa con el mejor ajuste a nuestros datos. No obstante, nos encontramos con un "Suggested power trasnformation" = 1.87. Este valor es la transformación de potencia para estabilizar la dispersión y se calcula como la pendiente de la línea ajustada al gráfico. Si hay homocedasticidad en el modelo el valor debe ser próximo a uno. Como comprobamos está un poco alejado del 1. No obstante, según la prueba de homocedasticidad anterior, nos mostraba que la varianza era constante. A pesar de esta pequeña contradicción asumimos que la varianza sigue siendo constante en nuestro modelo acorde a la prueba de homocedasticidad realizada con anterioridad.

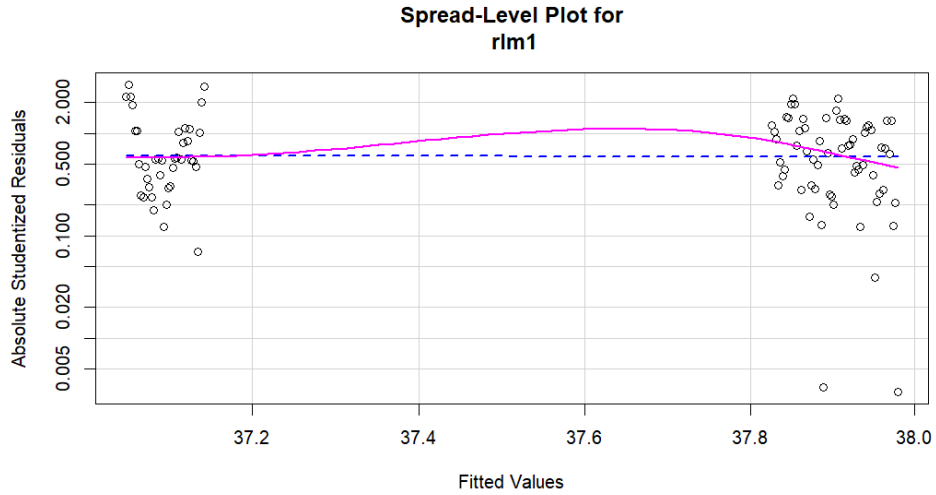

Figura 11: Gráfico de dispersión de los residuos estandarizados absolutos del modelo de regresión lineal múltiple frente a los valores ajustados. Se muestran dos conjuntos de datos agrupados, correspondiéndose a los dos grupos formados en la actividad física. Suggested power transformation = 1.87.

A continuación, analizaremos los datos en busca de multicolinealidad, este fenómeno se presenta en el análisis de regresión cuando dos o más variables predictoras están altamente correlacionadas entre sí. Es por eso por lo que el estudio de la colinealidad es extremadamente importante en este contexto, para evitar y predecir los problemas de multicolinealidad. Para detectar la multicolinealidad utilizaremos el factor de inflación de la varianza (FIV). Para hallar el FIV se utilizó la función *vif()* en R, y posteriormente con una función booleana utilizando el 2 como valor de referencia se confirmó que nuestros datos no presentan multicolinealidad.

3.2.2. Predicciones.

En esta ocasión se decidió poner aprueba el modelo y comparar los resultados predichos con los resultados reales extraídos del estudio realizado por Reynolds. Se escogieron al azar 4 horas diferentes, mientras que el estado físico se obtendrá en base a la actividad física más probable de los castores a esas horas.

- Primer caso: 01:00h:
    - Valor real: 37.75

    - Valor predicho: 38.20

En este primer caso nos encontramos en mitad de la noche donde los castores están ejerciendo una intensa actividad física. Al comparar los valores vemos que el valor predicho está muy por encima del valor real, pudiendo ser en este caso erróneo.

- Segundo caso: 12:00h:
    - Valor real: 37.01
    - Valor predicho: 37.10

En el segundo caso, el castor se encuentra descasando en el refugio, y como vemos en valor real y el predicho están bastante próximos.

- Tercer caso: 17:30h:
    - Valor real: 37.91
    - Valor predicho: 37.95

A continuación, observamos que a las 17:30 es una hora clave, donde los castores salen de su refugio para realizar sus actividades nocturnas. Vemos nuevamente como los valores están muy próximos entre sí.

- Cuarto caso: 19:30:
    - Valor real: 37.74
    - Valor predicho: 37.92

En el último caso, comprobamos como el valor real y el predicho también están bastante próximos, aunque algo más alejados que los dos últimos casos.

4. Discusión.

En general, con el objetivo de prever la temperatura corporal de los castores mediante un sólido modelo de regresión lineal, se ha indagado en la intersección entre la teoría estadística y la biología aplicada. A lo largo de este estudio, se ha cumplido meticulosamente con las hipótesis fundamentales del modelo, a excepción de la independencia, estableciendo relaciones significativas entre múltiples variables predictoras y la temperatura corporal de los castores.

Como se ha comprobado anteriormente, a pesar de la robustez del modelo y su concordancia con las hipótesis estadísticas, nos encontramos con valores ambiguos y en ocasiones alejados del valor verdadero. Estos valores nos presentan la complejidad inherente de los sistemas biológicos y la interconexión de factores que desafían nuestra capacidad de modelar con perfección.

La elección de un modelo de regresión se justifica por su capacidad para capturar y cuantificar relaciones complejas entre variables, permitiendo así una aproximación cuantitativa y sistemática a la predicción de la temperatura corporal. Con este informe se ha desarrollado un modelo de predicción robusto que cumple todas las hipótesis básicas del modelo de regresión, a excepción de la independencia, dado que tanto las pruebas de aleatoriedad como la visualización de los datos indican que no siguen un patrón aleatorio. La existencia de autocorrelación de primer orden en los residuos conlleva a la violación

de la independencia del modelo y a una ineficiencia de los estimadores. Existen alternativas que permiten abordar la falta de independencia de los errores e incluyen el uso de modelos de regresión robustos, como los modelos de ecuaciones generalizadas (GEE) en el caso de datos correlacionados [8], que no son objeto de estudio en este informe, pero que podrían plantear una solución a la falta de independencia de nuestro modelo. En cambio, los datos muestran una tendencia de agrupamiento en dos conjuntos, que se encuentran en los 36.8ºC (correspondiéndose a una actividad física nula y en el tiempo desde las 9h hasta las 16h) y en los 37.9ºC, (correspondiéndose a una actividad física intensa y en el tiempo desde las 16h hasta las 2h). El resultado del agrupamiento de los datos que observamos en los distintos gráficos es el resultado de la presencia de variables categóricas binarias como la actividad física.

No obstante, hay que tener en cuenta que, al formular las distintas hipótesis, el nivel de significación es del 10% (o $\alpha = 0.10$), haciendo que el modelo de regresión presente un 10% de probabilidad de cometer un error de tipo I, que consiste en rechazar incorrectamente una hipótesis nula cuando es verdadera. Al aceptar este nivel de significación implica un compromiso entre ser más tolerantes a los errores de tipo I y aceptar una menor capacidad para detectar efectos reales. En este caso la elección del nivel de significación se consideró a efectos prácticos a la hora de calcular la temperatura corporal de los castores con un valor aproximado, y, por tanto, que se presente un error de tipo I con una probabilidad del 10% no es tan relevante, pero siempre hay que tenerlo en cuenta.

En este estudio, se ha desarrollado un modelo de predicción para la temperatura corporal de los castores utilizando la regresión lineal múltiple, pero la integración de un modelo no lineal basado en el aprendizaje automático podría proporcionar mejoras sustanciales [9].

En primer lugar, es fundamental reconocer los desafíos inherentes a la medición de la temperatura en animales que se desplazan libremente en su entorno natural. Las limitaciones de los métodos convencionales, como la termometría, se hacen evidentes, ya que la obtención de datos continuos y precisos puede resultar complicada debido a los movimientos erráticos y la variabilidad ambiental. La implementación de un modelo no lineal, respaldado por técnicas de aprendizaje automático, ofrece una solución prometedora. Los modelos no lineales tienen la capacidad de capturar relaciones más complejas y no lineales en los datos, lo que podría ser crucial al abordar la variabilidad en los patrones de la temperatura corporal de los castores [8]. Al entrenar el modelo con datos históricos y correlacionar múltiples variables, se pueden obtener predicciones más precisas y contextualmente relevantes. Un estudio publicado por el doctor Guifeg Jia, permite evaluar la temperatura de los cerdos sin contacto e incorporando factores ambientales [8], generando un modelo de predicción no lineal basado en el aprendizaje automático, lo que podría ser interesante aplicar a este problema para hallar la temperatura de los castores.

Además, el aprendizaje automático permite que el modelo se adapte y evolucione a medida que se acumulan más datos. Esto es especialmente beneficioso en entornos dinámicos donde los patrones de comportamiento y las respuestas fisiológicas pueden cambiar con el tiempo y las estaciones. Entre estos algoritmos de aprendizaje automático

se destacan Support Vector Machine, Random Forest, y Back Propagation Neural Network, que se aplican ampliamente en problemas prácticos que implican reconocimiento de patrones y regresión en diversos campos [8].

5. Conclusión.

En conclusión, pese a que el modelo presente algunas limitaciones a la hora de poder calcular la temperatura corporal del *Castor canadensis*. Algunos de los valores contradictorios y limitaciones pueden derivar de fenómenos o variables no contempladas en este modelo, como puede ser la edad, el peso, altitud, humedad, etc. Que pueden variar enormemente la temperatura corporal. Los resultados de este informe permiten una reflexión más profunda y una exploración de factores adicionales que podrían influir en la temperatura corporal de los castores.

La falta de cumplimiento de la hipótesis de aleatoriedad en nuestro modelo de regresión lineal múltiple podría plantear preocupaciones sobre la validez de las inferencias estadísticas. Sin embargo, a pesar de esta falta de aleatoriedad, el modelo cumple con la mayoría de las hipótesis básicas de un modelo de regresión lineal múltiple, siendo este un modelo eficiente, que podría indicar que las estimaciones de los parámetros aún son precisas y que el modelo puede proporcionar buenas predicciones, aunque con precaución en la interpretación de los resultados.

El propósito final es proporcionar una herramienta eficaz para investigadores y conservacionistas, permitiéndoles anticipar y comprender mejor las respuestas térmicas de los castores en distintos entornos. En este caso, existe un desafío constante de mejorar la precisión del modelo para adaptarse a la complejidad de los sistemas biológicos. No obstante, nuestro modelo de regresión lineal se proyecta como una valiosa herramienta para comprender y prever la temperatura corporal de los castores, contribuyendo así al conocimiento de la ecología y biología aplicada.

6. Bibliografía.

[1] Saltz, D., & White, G. C. (2013). Wildlife Management. En Elsevier eBooks (pp. 403-407). https://doi.org/10.1016/b978-0-12-384719-5.00274-4

[2] Castor canadensis (American beaver). (s. f.). Animal Diversity Web. https://animaldiversity.org/accounts/Castor_canadensis/

[3] U, A. B., A, J. O., & Ramírez, M. (2008). Impacto del castor (Castor canadensis, rodentia) en bosques de Lenga (Nothofagus pumilio) de Tierra del Fuego, Chile. Bosque, 29(2). https://doi.org/10.4067/s0717-92002008000200009

[4] McCafferty, D. J., Gallon, S. L., & Nord, A. (2015). Challenges of measuring body temperatures of free-ranging birds and mammals. Animal Biotelemetry, 3(1). https://doi.org/10.1186/s40317-015-0075-2

[5] Taylor, N. A., Tipton, M. J., & Kenny, G. P. (2014). Considerations for the measurement of core, skin and mean body temperatures. Journal of Thermal Biology, 46, 72-101. https://doi.org/10.1016/j.jtherbio.2014.10.006

[6] *R: Body temperature series of two beavers*. (s. f.-b). https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/beavers.html

[7] Playground, P. (2015). *Body Temperature Series of two beavers | R Bloggers*. R-bloggers. https://www.r-bloggers.com/2015/12/body-temperature series-of-two-beavers/

[8] Harrell, F. E. (2001). Regression modeling strategies. En *Springer series in statistics*. https://doi.org/10.1007/978-1-4757-3462-1

[9] Jia, G., Li, W., Meng, J., Tan, H., & Feng, Y. (2020). Non-Contact evaluation of pigs' body temperature incorporating environmental factors. *Sensors, 20*(15), 4282. https://doi.org/10.3390/s20154282