Regresión, modelos y métodos
Prueba de evaluación continua 1

Susana Barcelo, Víctor J. Gallardo,
Geòrgia Escaramís, Santiago Ríos y Francesc Carmona

Fecha publicación del enunciado: 22-04-2023
Fecha límite de entrega de la solución: 7-05-2023

Presentación Esta PEC consta de ejercicios similares a los planteados en los ejercicios con los que podréis contrastar vuestra asimilación de los conceptos y métodos presentados en las tres últimas unidades.

Objetivos El objetivo de esta PEC es trabajar los conceptos de regresión múltiple trabajados en la primera parte de la asignatura.

Descripción de la PEC Debéis responder cada problema por separado. Recordad que tan importante como el resultado es el razonamiento y el proceso que os lleva a ello, es decir el consultor debe poder ver no tan sólo donde habéis llegado sino también como y porqué habéis llegado hasta allí. Incluid el código de R en la solución.

Criterios de valoración Cada PEC representa un 50 % de la nota de la asignatura. La presentación de los ejercicios aportará una puntuación que **se sumará** a los puntos obtenidos por las PECs.
Se valorará positivamente la contención en las respuestas del software y negativamente los volcados de datos innecesarios.

Código de honor Cuando presentáis ejercicios individuales os adherís al código de honor de la UOC, con el que os comprometéis a no compartir vuestro trabajo con otros compañeros o a solicitar de su parte que ellos lo hagan. Asimismo aceptáis que, de proceder así, es decir, en caso de copia probada, la calificación total de la PEC será de cero, independientemente del papel (copiado o copiador) o la cantidad (un ejercicio o todos) de copia detectada.

Formato Para hacer la entrega se tiene que enviar un mensaje al buzón de entregas del aula. En este mensaje debéis adjuntar un fichero PDF (obtenido a partir de vuestra solución en Word, Open Office, LaTeX, LyX o RMarkdown). El nombre del fichero debe ser la composición de vuestro apellido y vuestro nombre seguido de `_Reg_PEC1.pdf` (por ejemplo: si vuestro nombre es "Josep Benet", el fichero debe llamarse `benet_josep_Reg_PEC1.pdf`). También puede ser en formato HTML.

*Es **importante** que el examen sea legible y, a ser posible, elegante. Como si fuera un informe a vuestro jefe. Por ello valoraremos que separéis el código **R** (no necesario para la comprensión de la resolución) de los resultados y la discusión. Podéis hacerlo por ejemplo dejando el código completo en un apéndice. En medio de las explicaciones podéis poner vuestro código pero controlad la longitud de los resultados (evitad por ejemplo páginas enteras que únicamente contienen números).*

Ejercicio 1 (35 pt.)

Un grupo de científicos norteamericanos están interesados en encontrar un hábitat adecuado para reintroducir una especie rara de escarabajos tigre, llamada *cicindela dorsalis dorsalis*, los cuales viven en playas de arena de la costa del Atlántico Norte. Se muestrearon 12 playas y se midió la densidad de estos escarabajos tigre. Adicionalmente se midieron una serie de factores bióticos y abióticos tales como la exposición a las olas, tamaño de la partícula de arena, pendiente de la playa y densidad de los anfípodos depredadores.

Los datos se hallan en la hoja de cálculo `cicindela.xlsx`.

Cicindela dorsalis.

(a) Ajustar un modelo de regresión lineal múltiple que estime todos los coeficientes de regresión parciales referentes a todas las variables regresoras y el intercepto.

¿Es significativo el modelo obtenido? ¿Qué test estadístico se emplea para contestar a esta pregunta. Plantear la hipótesis nula y la alternativa del test.

¿Qué variables han salido significativas para un nivel de significación $\alpha = 0.10$?

(b) Calcular los intervalos de confianza al 90 y 95 % para el parámetro que acompaña a la variable `AmphipodDensity`. Utilizando sólo estos intervalos, ¿qué podríamos haber deducido sobre el p-valor para la densidad de los anfípodos depredadores en el resumen del modelo de regresión? ¿Qué interpretación práctica tiene este parámetro β_4?

(c) Estudiar la posible multicolinealidad del modelo con todas las regresoras calculando los VIFs.

(d) Considerar el modelo más reducido que no incluye las variables exposición a las olas y la pendiente de la playa y decidir si nos podemos quedar con este modelo reducido mediante un contraste de modelos con el test F para un $\alpha = 0.05$. Escribir en forma paramétrica las hipótesis H_0 y H_1 de este contraste. Comparar el ajuste de ambos modelos.

(e) Calcular y dibujar una región de confianza conjunta al 95 % para los parámetros asociados con `Sandparticlesize` y `AmphipodDensity` con el modelo que resulta del apartado anterior.

Dibujar el origen de coordenadas. La ubicación del origen respecto a la región de confianza nos indica el resultado de una determinada prueba de hipótesis. Enunciar dicha prueba y su resultado.

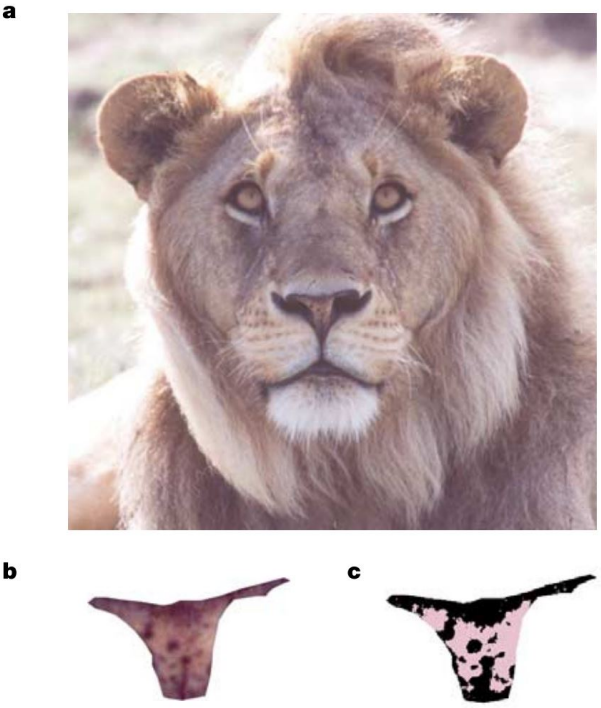
(f) Con el modelo reducido del apartado (d), predecir en forma de intervalo de confianza al 95 % la densidad de los escarabajos tigre previsible para una playa cercana a un conocido hotel donde el tamaño de partícula de arena es 5 y la densidad de anfípodos depredadores es 11. Comprobar previamente que los valores observados no suponen una extrapolación.

Ejercicio 2 (35 pt.)

En el trabajo de Whitman et al. (2004) se estudia, entre otras cosas, la relación entre la edad de los leones y la proporción oscura en la coloración de sus narices. En el archivo `lions.csv` disponemos de los datos de 105 leones machos y hembras de dos áreas de Tanzania, el parque nacional de Serengueti y el cráter del Ngorongoro, entre 1999 y 2002. Las variables registradas son la edad conocida de cada animal y la proporción oscura de su nariz a partir de fotografías tratadas digitalmente (ver figura adjunta).

En la figura 1 se reproduce el gráfico de dispersión de la figura 4 del artículo con el cambio de coloración de la nariz según la edad de machos y hembras en las dos poblaciones separadas.

Nota: Los datos se han extraído principalmente del gráfico del artículo de Whitman et al. (2004) y por lo tanto son aproximados. Algunos paquetes de R contienen un *data.frame* con una parte de estos datos. Por ejemplo `LionNoses` del paquete `abd` contiene los datos de todos los machos. En consecuencia, los resultados numéricos de vuestro análisis pueden ser ligeramente distintos a los del trabajo original.



Fotografía de un león macho y digitalización de su nariz.

(a) Reproducir el gráfico de dispersión de la figura 1 (figura 4d del artículo) lo más fielmente posible al original, ya que se trata de una exigencia de los editores de la revista.

(b) En el artículo se destacan los siguientes resultados:

> After controlling for age, there was no effect of sex on nose colour in the Serengeti, but Ngorongoro males had lighter noses than Ngorongoro females.

Ajustar un primer modelo sin considerar la posible interacción entre el sexo y las áreas y contrastar si el sexo es significativo en el modelo así ajustado y en los modelos separados según el área.

(c) Otro resultado destacado es que para los machos hay diferencias según el área. Contrastar este resultado y dibujar las rectas de regresión para las dos áreas que se obtienen del modelo.

(d) En la tabla 1 del artículo de Whitman et al. se dan los intervalos de confianza al 95 %, al 75 % y al 50 % para predecir la edad de una leona de 10 años o menos según su proporción de pigmentación oscura en la nariz. La primera cuestión es: ¿sirven para esto los modelos estudiados en los apartados anteriores?

Reproducir la fila de la tabla 1 para una proporción del 0.50 según el modelo que proponen en el artículo.

Aclarar un detalle: lo que en la tabla 1 del artículo se llama s.e., standard error ¿qué es exactamente?

Nota: Recordemos también aquí que los resultados pueden ser ligeramente distintos a los del artículo por la utilización de datos aproximados.

Table 1 **Statistical relationship between nose blackness and age**

Proportion black	Estimated age in years (s.e.)	95% p.i.	75% p.i.	50% p.i.
0.10	2.66 (1.24)	0.17–5.15	1.21–4.10	1.81–3.50
0.20	3.25 (1.24)	0.77–5.72	1.81–4.69	2.41–4.09
0.30	3.84 (1.23)	1.37–6.30	2.40–5.27	3.00–4.68
0.40	4.42 (1.23)	1.97–6.89	3.00–5.86	3.59–5.26
0.50	5.02 (1.23)	2.56–7.48	3.59–6.45	4.18–5.85
0.60	5.61 (1.23)	3.14–8.07	4.18–7.04	4.77–6.44
0.70	6.20 (1.23)	3.73–8.66	4.77–7.63	5.36–7.04
0.80	6.79 (1.24)	4.32–9.26	5.35–8.23	5.95–7.63
0.90	7.38 (1.24)	4.90–9.87	5.94–8.82	6.54–8.22
1.00	7.97 (1.25)	5.58–10.47	6.52–9.42	7.12–8.82

s.e., standard error; p.i., predicted interval. Predicted values are based upon the least-squares regression of a truncated data set for 63 known-aged females in the Serengeti and Ngorongoro aged ≤10 yr ($y = 2.0667 + 5.9037\arcsin(x)$; $r^2 = 0.75$, $P < 0.0001$). Predicted intervals at 95%, 75% and 50% are included for upper and lower bounds.

Tabla 1: Tabla 1 del artículo de Whitman et al.

Ejercicio 3 (30 pt.)

(a) Verificar las hipótesis de Gauss-Markov y la normalidad de los residuos del modelo completo del apartado (b) del ejercicio 2. Realizar una completa diagnosis del modelo para ver si se cumplen las condiciones del modelo de regresión: normalidad, homocedasticidad,... y estudiar la presencia de valores atípicos, de alto *leverage* y/o puntos influyentes.

Construir los gráficos correspondientes y justificar su interpretación. ¿Podemos considerar el modelo ajustado como fiable?

(b) Teniendo en cuenta que la variable respuesta de la regresión del apartado (b) del ejercicio 2 es una proporción, ¿presenta algún problema este modelo? ¿Qué alternativas nos podemos plantear para mejorar el ajuste de los datos?

Atendiendo a la naturaleza de la variable respuesta, ¿hay alguna transformación adecuada?

(c) Aplicar la transformación más adecuada a la variable respuesta del modelo considerado. Comparar los dos modelos: con y sin la transformación. ¿Qué modelo es mejor? Justificar la respuesta.

(d) Realizar una rápida diagnosis del modelo transformado. ¿Estamos satisfechos con este nuevo modelo? ¿Qué otro ajuste nos podemos plantear para mejorar el modelo?

(e) Discutir la utilización de la transformación arcoseno en el modelo del apartado (d) del ejercicio 2.

Referencias

[1] Whitman, K., A.M. Starfield, H.S. Quadling and C. Packer. 2004. Sustainable trophy hunting of African lions. *Nature* 428: 175-178.

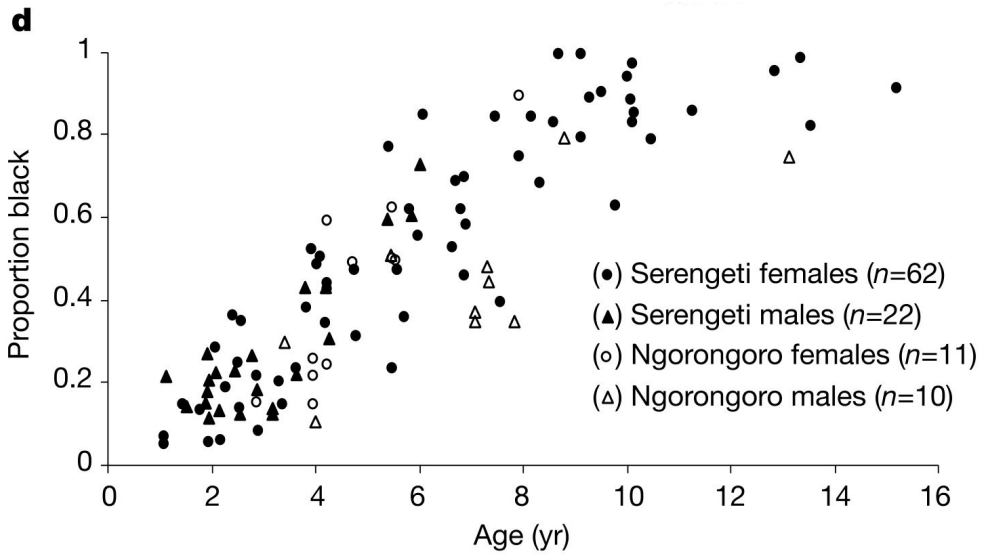


Figura 1: Cambio de coloración de la nariz según la edad de machos y hembras en dos poblaciones separadas (Figura 4d del artículo de Whitman et al. (2004)).