



## Regresión, modelos y métodos Prueba de evaluación continua 2

Francesc Carmona y Mireia Besalú

Fecha publicación del enunciado: 4-1-2020

Fecha límite de entrega de la solución: 19-1-2020

**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 segunda 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_PEC2.pdf` (por ejemplo: si vuestro nombre es "Jordi Pujol", el fichero debe llamarse `pujol_jordi_Reg_PEC2.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** de los resultados y la discusión. Podéis hacerlo por ejemplo dejando el código completo en un apéndice -o en un archivo `.R` adjunto-. 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 (25 pt.) Mecanismo neural de la nocicepción.

El gusano *Caenorhabditis elegans* de la familia *Rhabditidae* mide aproximadamente 1 mm de longitud y vive en ambientes templados. Es un importante modelo de estudio en biología, muy especialmente en genética del desarrollo desde los años 70, en parte debido a su pequeño número de neuronas y su genoma fácilmente manipulable. La nocicepción es la percepción neural de un estímulo real o potencialmente dañino. En *C. elegans* evoca un comportamiento de abstinencia autoconservante. Sin embargo, la estimulación repetida puede derivar en una respuesta de abstinencia o habituación reducida.



Los investigadores compararon la respuesta de retirada a estímulos de luz perturbadores en *C. elegans* de tipo salvaje y una línea mutante de *C. elegans* que exhibe una respuesta más lenta de las neuronas sensoriales PVD. La falta de reacción indica habituación. Aquí están los porcentajes de animales probados que exhibieron una reacción de abstinencia a un estímulo nocivo que consiste en un número variable de pulsos de luz consecutivos (IndMutant = 1 para la línea mutante y 0 para el tipo salvaje):

% Reacting	Pulses	IndMutant	% Reacting	Pulses	IndMutant	% Reacting	Pulses	IndMutant
93	1	1	54	15	1	91	9	0
81	2	1	56	16	1	82	10	0
73	3	1	50	17	1	86	11	0
64	4	1	42	18	1	82	12	0
73	5	1	42	19	1	91	13	0
68	6	1	29	20	1	85	14	0
64	7	1	96	1	0	83	15	0
65	8	1	100	2	0	70	16	0
50	9	1	92	3	0	74	17	0
68	10	1	96	4	0	64	18	0
57	11	1	92	5	0	68	19	0
50	12	1	92	6	0	78	20	0
62	13	1	84	7	0			
44	14	1	96	8	0			

- (a) En una primera aproximación para comparar los porcentajes de reacción de los gusanos mutantes con los salvajes, nos planteamos hacer un test de comparación de dos muestras independientes y un gráfico adecuado, sin tener en cuenta los pulsos. Para ello debemos observar que la variable respuesta es un porcentaje y el test  $t$  de Student habitual no sirve.

Entre las posibles soluciones tenemos las siguientes:

1. Aplicar un test de Welch a los datos transformados con la función normalizante  $\arcsin(\sqrt{p})$ .
2. Hacer un test no paramétrico como el de Mann-Whitney-Wilcoxon.
3. Aplicar un test de permutaciones para comparar las dos muestras.
4. Realizar una regresión logística binomial.
5. Calcular una *regresión beta* que es especialmente apropiada para proporciones.

Las dos últimas propuestas son las más acertadas, ya que las otras contemplan comparar medias o medianas de porcentajes. En todo caso, realizar *tres de las cinco soluciones con una de ellas entre las dos últimas* y comentar su resultado.

*Nota 1:* Para aplicar un test de permutaciones se puede recurrir al paquete `coin` o al paquete `lmPerm` de **R**.

*Nota 2:* Para saber sobre la regresión beta y su aplicación con **R** podemos consultar:

[https://rcompanion.org/handbook/J\\_02.html](https://rcompanion.org/handbook/J_02.html)

- (b) Dibujar el gráfico de dispersión del porcentaje de reacción en función del número de pulsos de luz con dos rectas que representen las dos subpoblaciones por separado, sin modelo común. Describir la forma de la relación. ¿Qué sugiere sobre el patrón de respuestas de abstinencia en las dos subpoblaciones de *C. elegans* para un número creciente de pulsos de luz? ¿Cómo encaja su respuesta en el contexto de habituación? Explique por qué estos resultados sugieren una participación de las neuronas sensoriales de PVD en la nocicepción y la habituación.

*Nota:* En este apartado consideramos la variable respuesta el porcentaje de reacción sin ninguna transformación.

- (c) Proponer un modelo lineal que permita estudiar la asociación de la variable respuesta porcentaje de reacción (sin ninguna transformación) con la variable **Pulses** y el factor **IndMutant**.

¿Cuál es la ecuación del modelo final propuesto? ¿Es un modelo significativo? ¿Qué tanto por ciento de la variabilidad del porcentaje de reacción queda explicado por el modelo? ¿Todos los coeficientes de las variables explicativas del modelo son significativos?

Dibujar el gráfico de dispersión del apartado anterior pero con las rectas resultantes del modelo.

*Nota:* En este apartado no haremos ninguna transformación de la variable respuesta, ni utilizaremos ningún modelo generalizado, ya que confiamos en la validez del modelo balanceado.

- (d) Ahora contestaremos las preguntas y dibujaremos el gráfico del apartado anterior si realizamos la transformación normalizante  $\arcsin(\sqrt{p})$  sobre la variable respuesta.

Hallar el intervalo de confianza al 95% para la media del porcentaje de mutantes que reaccionan a 10 pulsos de luz en condiciones experimentales parecidas.

*Nota:* Habrá que deshacer la transformación para dibujar “las rectas”.

## Ejercicio 2 (50 pt.)

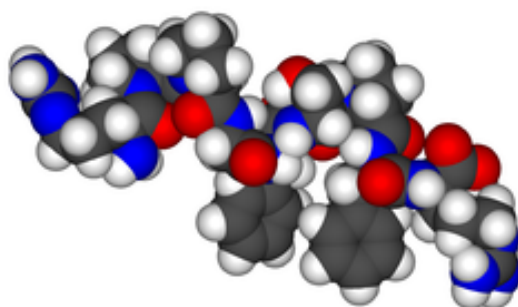
En este ejercicio se estudian diferentes formas de examinar un modelo de regresión múltiple. Los datos de Umetrics (1995)[4] provienen del campo del descubrimiento de fármacos. Se desarrollan nuevos medicamentos a partir de productos químicos que son biológicamente activos. Probar la actividad biológica de un compuesto es un procedimiento costoso, por lo que es útil poder predecir la actividad biológica a partir de mediciones químicas más baratas.

De hecho, la química computacional hace posible calcular ciertas mediciones químicas sin siquiera hacer el compuesto. Estas medidas incluyen el tamaño, la lipofilia y la polaridad en varios lugares de la molécula.

Los datos[3] se hallan en el archivo `penta.dat`. Mejor si eliminamos las observaciones con algún valor perdido.

Queremos estudiar la relación entre estas mediciones y la actividad del compuesto, representada por el logaritmo de la actividad relativa de bradiquinina  $\log.RAI$ . Debemos tener en cuenta que estos datos contienen muchos predictores en relación con el número de observaciones. Será útil hallar algunos factores

Bradiquinina



predictivos subyacentes que expliquen la mayor parte de la variación en la respuesta. Por lo general, el modelo se ajusta a parte de los datos (el conjunto de “entrenamiento” o “trabajo”) y la calidad del ajuste se juzga por lo bien que predice la otra parte de los datos (el conjunto de “prueba” o “predicción”). Para este ejercicio tomaremos los siguientes conjuntos:

```
set.seed(321)
idx <- sample(1:30, size = 20, replace = F)
ptrain <- penta[idx,]
ptest <- penta[-idx,]
```

Para evaluar el ajuste utilizaremos la raíz cuadrada de la media de los errores al cuadrado (RMSE).

*Nota:* Los mismos datos se hallan en el *data.frame* `Penta` del paquete `mvdalab` de **R**.

- (a) Comprobar con los factores de inflación de la varianza que el modelo lineal completo y con todos los datos padece un problema grave de multicolinealidad.

Esto justifica que debemos reducir el número de variables predictoras o utilizar algún método alternativo.

- (b) Estudiar el modelo reducido que proporciona el método *stepwise*.

Se trata de comparar los RMSE del modelo reducido y del modelo completo para el grupo de entrenamiento y para el grupo de prueba. También de ver si el modelo reducido salva el problema de multicolinealidad.

- (c) Hallar el mejor modelo por el criterio del AIC y por el  $R^2$  ajustado. ¿Coinciden?

¿Es el mismo modelo que selecciona el *stepwise*?

- (d) Estudiar una regresión por componentes principales con estos datos. ¿Qué tal funciona con 8 componentes para el grupo de entrenamiento?

¿Cual es el número de componentes que se recomienda si hacemos una validación cruzada? Con ese número mínimo de componentes, ¿cual es el RMSE para el conjunto de prueba?

- (e) Estudiar una regresión PLS. ¿Cuántas componentes selecciona la validación cruzada?

¿Cual es la variabilidad de las variables predictoras explicada por 3 componentes?

Con el paquete `plsdepot` se pueden estudiar las correlaciones entre las variables y las componentes y realizar el gráfico llamado *círculo de correlaciones*. ¿Cual es la variable predictora mejor correlacionada con la primera componente?

*Nota:* Recordemos que ya hemos fijado la semilla aleatoria al principio: `set.seed(321)`.

- (f) Estudiar también el método de Ridge Regression con la función `lm.ridge()`. Además de los valores RMSE para el grupo de entrenamiento y de prueba, debemos acompañar el análisis con un gráfico de los coeficientes.

Para hallar el lambda óptimo podemos empezar por un límite superior bajo y aumentar sucesivamente ese límite hasta que el valor óptimo no sea el límite superior.

Aunque siempre es mejor estandarizar los datos, en este ejercicio no lo haremos para no complicar más los cálculos del RMSE del conjunto de prueba. Internamente, lo hace la propia función `lm.ridge()`<sup>1</sup>.

*Nota1:* Si el resultado del ajuste de las predicciones a los datos es decepcionante podríamos optar por utilizar la función `cv.glmnet(..., alpha=0)` del paquete `glmnet`. Esta función no necesita que

---

<sup>1</sup>Se puede comprobar si ejecutamos la instrucción `MASS::lm.ridge` (sin paréntesis) en la consola de **R**.

le indiquemos una secuencia de valores de lambda, aunque sí requiere que le pasemos los datos como objeto `matrix`. Además podemos utilizar la validación *leave one out* simplemente haciendo que el número de carpetas (*folds*) sea exactamente el número de observaciones.

*Nota2:* Aunque no hace falta, si se utilizan las funciones `cv.glmnet()` y `glmnet()` para el cálculo del modelo Ridge Regression, se puede comprobar que los resultados difieren notablemente de los que tenemos con la función `lm.ridge()`. Una sencilla explicación se puede leer en

<https://stats.stackexchange.com/questions/74206/ridge-regression-results-different-in-using-lm-ridge-and-glmnet>

(g) Finalmente estudiaremos el método LASSO con la función `lars()` del paquete del mismo nombre.

*Nota:* Tal vez el resultado es demasiado restrictivo y hay que aumentar el número de variables seleccionadas, lo que equivale a aumentar ligeramente el valor de lambda.

### Ejercicio 3 (25 pt.)

Es posible utilizar la regresión logística para estudiar la discriminación entre dos grupos dado un conjunto de variables explicativas. Para más detalles podéis consultar el apartado 8.10 del libro de Manly[1].



(a) Realizar un análisis discriminante con ayuda de la regresión logística con los datos de los 49 gorriones hembra después de una tormenta.

Reproducir con todo detalle la tabla 8.6 de la página 153 del libro de Manly.

(b) Contrastar con un test  $\chi^2$  la significación de la regresión y explicar su resultado.

(c) Realizar un gráfico como el de la figura 8.2 de la página 156 del libro de Manly pero con los datos de este ejercicio y valorar el resultado.

(d) Calcular un intervalo de confianza para el parámetro de la variable  $x_4 = \text{length humerus}$  y para su *odds ratio*.

(e) Calcular la tabla de confusión de la clasificación obtenida con la regresión logística.

## Referencias

- [1] Bryan F.J. Manly and Jorge A. Navarro Alberto (2017), *Multivariate Statistical Methods: A Primer*, Fourth Edition. Chapman and Hall/CRC. Springer-Verlag.
- [2] SAS/STAT(R) 9.22 User's Guide, "The PLS Procedure".
- [3] <https://v8doc.sas.com/sashtml/stat/chap51/sect19.htm>
- [4] Umetrics, Inc. (1995), *Multivariate Analysis (3-day course)*, Winchester, MA.