

Regresión, modelos y métodos Prueba de evaluación continua 2

Francesc Carmona y Mireia Besalú

Fecha publicación del enunciado: 6-6-2020
Fecha límite de entrega de la solución: 21-6-2020

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 (50 pt.)

El archivo `peru.txt` contiene algunas variables posiblemente relacionadas con la presión sanguínea de $n = 39$ peruanos que se han trasladado de las zonas rurales de gran altitud a las zonas urbanas de menor altitud. Considerar un modelo de regresión múltiple para predecir la presión sistólica Y a partir de las variables:

$$X_1 = \text{age}$$

$$X_2 = \text{years in urban area}$$

$$X_3 = X_2/X_1 = \text{fraction of life in urban area}$$

$$X_4 = \text{weight (kg)}$$

$$X_5 = \text{height (mm)}$$

$$X_6 = \text{chin skinfold}$$

$$X_7 = \text{forearm skinfold}$$

$$X_8 = \text{calf skinfold}$$

$$X_9 = \text{resting pulse rate}$$



Machu Picchu, Peru.

- (a) Estudiar la posible multicolinealidad de este modelo.
- (b) Eliminar una única observación de la muestra de forma que el modelo mejore apreciablemente. Razonar la elección.
- (c) Con los 38 datos restantes, hallar el “mejor” modelo consensuado por dos métodos diferentes de selección de variables como, por ejemplo, R^2_{adj} y C_p de Mallows.
 - (i) ¿Cuáles son las variables seleccionadas?
 - (ii) ¿Cuál es el coeficiente de determinación ajustado de este modelo? Compararlo con el del modelo completo.
 - (iii) ¿Se gana en eficiencia con el modelo reducido? Comparar los intervalos de confianza de la estimación del coeficiente de la variable *Age*.
- (d) Los investigadores sugieren adoptar el modelo reducido que contenga únicamente las variables significativas ($\alpha = 0.1$) con el test *t* en sustitución del modelo completo con las 9 variables explicativas. ¿Es ese un buen criterio de selección?
Realizar un test adecuado que resuelva su sugerencia. Discutir el resultado en consonancia con los resultados obtenidos en el apartado anterior.
- (e) Comprobar si hemos solucionado el problema de multicolinealidad en el modelo reducido del apartado anterior.
Como los investigadores no quieren prescindir de más variables, se plantea una regresión *Partial Least Squares* (PLS). ¿Cuántas componentes se necesitan para minimizar el RMSEP?
Calcular los coeficientes de las variables originales, también para β_0 , que proporciona este método con el número de componentes necesario.
¿Es adecuado este método de regresión con estas variables? ¿Es útil?
- (f) Siguiendo con el modelo reducido, otra posibilidad es utilizar la *Ridge Regression*. ¿Cuáles son los coeficientes obtenidos? Explicar brevemente las ventajas e inconvenientes de este método frente a la selección de variables.
Calcular el RMSE de la regresión OLS, PLS (con 5, 4, 3 y 2 componentes) y Ridge (con λ óptima por GCV) para el modelo reducido.

¿Cuál es la valoración con todo lo que sabemos hasta ahora?

- (g) Sabemos que el RMSE calculado en un modelo para todos los datos observados es muy optimista. Es mejor un cálculo por validación cruzada.

Con el modelo reducido de los apartados anteriores y para comparar los métodos estudiados (OLS, PLS (con 4 componentes) y Ridge (con λ óptimo por GCV) haremos lo siguiente:

1. Dividiremos los datos aleatoriamente en dos grupos, uno de 8 observaciones (grupo test) y otro del resto (grupo train). Recordemos que el número total de observaciones es ahora de 38.
2. Ajustaremos cada modelo con el grupo train y calcularemos el RMSE con el grupo test.
3. Repetiremos los pasos 1 y 2 mil veces.
4. Finalmente compararemos los resultados para cada modelo con algún estadístico y también gráficamente con las densidades de los RMSE.

¿Qué podemos decir?

- (h) Recordemos que en la regresión OLS sabemos que

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{P}\mathbf{Y}$$

donde \mathbf{P} es la matriz proyección (*hat matrix*). Los grados de libertad usados por la regresión son exactamente el número de columnas linealmente independientes en la matriz de diseño \mathbf{X} .

Además se sabe que

$$\text{rg}(\mathbf{X}) = \text{rg}(\mathbf{X}'\mathbf{X}) = \text{rg}(\mathbf{P}) = \text{traza}(\mathbf{P})$$

ya que la matriz \mathbf{P} es idempotente.

En particular, si \mathbf{X} es de rango máximo, entonces $\text{rg}(\mathbf{X}) = p$ y $\text{traza}(\mathbf{P}) = p$, con p = número de columnas de \mathbf{X} .

Del mismo modo, los grados de libertad en la Ridge regression se definen con la traza de la matriz $\mathbf{H}(\lambda)$:

$$\text{traza}(\mathbf{H}(\lambda)) = \text{traza}[\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}']$$

Por otra parte, se demuestra con relativa facilidad que

$$\text{traza}(\mathbf{H}(\lambda)) = \sum_{j=1}^p \frac{d_j^2}{d_j^2 + \lambda}$$

donde d_1, \dots, d_p son los valores singulares de la matriz de diseño \mathbf{X} y, por lo tanto, d_1^2, \dots, d_p^2 son los valores propios de $\mathbf{X}'\mathbf{X}$.

De este modo los grados de libertad de la regresión regularizada por un λ forman una función monótona decreciente para λ . Cuando $\lambda = 0$ y la matriz de diseño es de rango máximo, los grados de libertad también son máximos con valor p . Por otra parte,

$$\lim_{\lambda \rightarrow \infty} \text{traza}(\mathbf{H}(\lambda)) = 0$$

En resumen, tenemos una forma de calcular el *rango efectivo* de la matriz de diseño asociada a la Ridge regression para un valor de λ .

Calcular los grados de libertad de la Ridge regression para el λ óptimo del apartado (e).

Effect of supplemental ascorbate on survival time of cancer patients.

Stomach Cancer			Bronchus Cancer			Colon Cancer		
Age	Days	Cont.	Age	Days	Cont.	Age	Days	Cont.
Females:			Females:			Females:		
61	124	38	48	87	13	76	135	18
62	19	36	64	115	49	58	50	30
66	45	12	Males:			70	155	57
69	876	19	74	74	33	68	534	16
59	359	55	74	423	18	74	126	21
Males:			66	16	20	76	365	42
69	12	18	52	450	58	56	911	40
63	257	64	70	50	38	74	366	28
79	23	20	77	50	24	60	99	28
76	128	13	71	113	18	Males:		
54	46	51	70	857	18	49	189	65
62	90	10	39	38	34	69	1,267	17
46	123	52	70	156	20	50	502	25
57	310	28	70	27	27	66	90	17
			55	218	32	65	743	14
			74	138	27	58	156	31
			69	39	39	77	20	33
			73	231	65	38	274	80

Tabla 1: Datos de Cameron and Pauling[1] para tres tipos de cáncer.

Ejercicio 2 (30 pt.)

En el trabajo de Cameron and Pauling[1] se presenta un estudio de los tiempos de supervivencia de 100 pacientes de cáncer terminal a los que se les administró un suplemento de ascorbato de sodio, vitamina C, como parte de su tratamiento rutinario y 1000 controles emparejados, pacientes similares que habían recibido el mismo tratamiento excepto por el ascorbato. El objetivo de la investigación fue determinar si el ascorbato de sodio suplementario prolongaba los tiempos de supervivencia de los pacientes con cáncer humano terminal.

Los datos se hallan en el archivo **Table 33.1** de la página

<https://www2.stat.duke.edu/courses/Spring01/sta114/data/andrews.html>

En el archivo descargado observaremos los 100 casos de la Tabla 1 del trabajo de Cameron and Pauling[1]. Las columnas de este archivo, a parte de las tres primeras, son las mismas que en la Tabla 1 del artículo. Falta añadir el tipo de cáncer y eliminar el símbolo + que indica una supervivencia superior al final del periodo de estudio.

En la tabla 1 se ven los datos del trabajo de Cameron and Pauling[1] sólo para tres tipos de cáncer: de estómago, de bronquios y de colon. Las variables en esta tabla se corresponden con la Tabla 1 de Cameron and Pauling así: $Age = A$, $Days = C$, $Cont. = D$.

- (a) Estudiar la transformación que mejora la distribución de los datos C y los datos D (100 observaciones en cada caso). Se puede utilizar el método de Box-Cox.

Una vez transformados, comparar si el tiempo de supervivencia C es superior al de los controles D con todas las observaciones.

- (b) Ahora estamos interesados en comparar la mejora en función del tipo de cáncer. Nos centraremos exclusivamente en los tres tipos de cáncer de la tabla 1 de más arriba y no tendremos en cuenta el sexo.

TABLE 9.1. Relationship between the conventional analysis of variance and ordinary least squares regression computations for the completely random experimental design.

Source of Variation	d.f.	Traditional AOV SS	Regression SS
Total uncorr	rt	$\sum \sum Y_{ij}^2$	$\mathbf{Y}'\mathbf{Y}$
Model	t	$\sum(Y_{i.})^2/r$	$\hat{\beta}'\mathbf{X}'\mathbf{Y}$
C.F.	1	$n\bar{Y}^2$	$n\bar{Y}^2$
Treatments	$t - 1$	$\sum(Y_{i.})^2/r - n\bar{Y}^2$	$\hat{\beta}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2$
Residual	$t(r - 1)$	$\sum \sum Y_{ij}^2 - \sum(Y_{i.})^2/r$	$\mathbf{Y}'\mathbf{Y} - \text{SS(Model)}$

Tabla 2: Tabla de las sumas de cuadrados para el ANOVA de un factor.

Consideremos la matriz de diseño \mathbf{X} correspondiente al modelo

$$y_{ij} = \mu_i + \epsilon_{ij} \quad \text{con } i = 1, 2, 3$$

donde no hay media común (o término de intercepción).

En el libro de regresión aplicada de Rawlings et al.[2], se muestra la tabla 2.

Calcular los elementos de dicha tabla con la matriz de diseño \mathbf{X} de este modelo y resolver con ellos el contraste $H_0 : \mu_1 = \mu_2 = \mu_3$ cuando la variable respuesta Y es el logaritmo de la razón entre la supervivencia de los tratados y la supervivencia de sus controles. ¿Cuál es la conclusión?

Nota: Habrá que tener en cuenta que en la tabla 2 se supone que el número de réplicas r es el mismo para todos los niveles, cosa que no pasa en este caso.

- (c) La edad de los pacientes presenta una cierta variabilidad y puede influir en su supervivencia.
Añadir a la matriz \mathbf{X} del apartado anterior el vector columna con las edades centradas.
Utilizar las sumas de cuadrados de los residuos de este modelo y del anterior para contrastar la importancia de ajustar con la edad.
¿Se puede utilizar un test t de Student?
- (d) Aunque la regresión de la edad en el modelo anterior pudiera no ser importante, se decidió que cada grupo debería tener su propia regresión sobre la edad para verificar si la edad no es importante en ninguno de los grupos.
Modificar adecuadamente la matriz de diseño para acomodar esta nueva situación y completar el test para la hipótesis nula de que la regresión sobre la edad es la misma en los tres grupos de cáncer.
¿Cuál es la conclusión?

Ejercicio 3 (20 pt.)

El conjunto de datos adjunto `diabetes.txt` es originario del *National Institute of Diabetes and Digestive and Kidney Diseases*. El objetivo del conjunto de datos es predecir si un paciente tiene o no diabetes, basándose en ciertas medidas diagnósticas incluidas en el conjunto de datos. Se pusieron varias limitaciones para la selección de estos casos de una base de datos más amplia. En particular, todos los pacientes aquí son mujeres de al menos 21 años de edad de herencia india Pima.

Los datos se podían obtener hasta hace poco en el dataset Pima Indian Diabetes 2 del *UCI Repository of machine learning databases* (Newman et al. 1998). Sin embargo, parece que ya no es accesible por problemas de permisos.

Un dataset similar se puede ver en

<https://www.kaggle.com/uciml/pima-indians-diabetes-database>

En el archivo `diabetes.txt` vamos a encontrar las siguientes variables:

<code>pregnant</code>	Number of times pregnant
<code>glucose</code>	Plasma glucose concentration (glucose tolerance test)
<code>pressure</code>	Diastolic blood pressure (mm Hg)
<code>triceps</code>	Triceps skin fold thickness (mm)
<code>insulin</code>	2-Hour serum insulin (mu U/ml)
<code>mass</code>	Body mass index (weight in kg/(height in m) ²)
<code>pedigree</code>	Diabetes pedigree function
<code>age</code>	Age (years)
<code>diabetes</code>	diabetes case (pos/neg)

donde la variable de interés es `diabetes`.

- (a) Ajustar un modelo de regresión logística para predecir la diabetes utilizando todas las otras variables como predictoras. Dar la ecuación del modelo obtenido y clasificar las variables según sean factores protectores o de riesgo para la diabetes.
- (b) Calcular el odds ratio de la variable `pedigree`, así como su intervalo de confianza.
- (c) Calcular el odds ratio y la probabilidad de tener diabetes para el individuo de la observación 9.
- (d) ¿Como valoras la bondad de ajuste del modelo? Realizar los contrastes o cálculos que se consideren necesarios.
- (e) Considerar ahora el modelo reducido con las variables `pregnant`, `glucose`, `mass`, `pedigree` y `age`. ¿Es significativa la variable `pregnant`?

Comparar los dos modelos.

Referencias

- [1] Cameron, E. and Pauling, L. (1987). Supplemental ascorbate in the supportive treatment of cancer: Re-evaluation of prolongation of survival times in terminal human cancer. *Proceedings of the National Academy of Science*, 75:4538-4542.
- [2] Rawlings, J.O., Pantula, S.G. and Dickey, D.A. (1998). *Applied Regression Analysis: A Research Tool*, Second Edition, Springer.