

DATOS PERSONALES		FIRMA
Nombre: <i>Diego Alberto</i>	DNI: <i>0922789151</i>	
Apellidos: <i>Coloma Borja</i>		

ESTUDIO	ASIGNATURA	CONVOCATORIA
MÁSTER UNIVERSITARIO EN INGENIERÍA MATEMÁTICA Y COMPUTACIÓN (PLAN 2016)	4391020006.- TÉCNICAS MULTIVARIANTES	Ordinaria Número periodo 1823
FECHA	MODELO	CIUDAD DEL EXAMEN
14-16/01/2022	Modelo - D	<i>Guayaquil</i>

Etiqueta Identificativa
-------------------------



\*02542412\*

Máster Universitario en Ingeniería Matemática y  
Computación (Plan 2016) | 1823

4391020006 - - Técnicas Multivariantes | 1823



## INSTRUCCIONES GENERALES

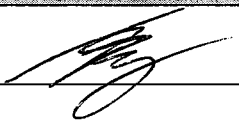
1. Ten disponible tu documentación oficial para identificarte, en el caso de que se te solicite.
2. Si tu examen consta de una parte tipo test, indica las respuestas en la plantilla según las características de este.
3. Debes contestar en el documento adjunto, respetando en todo momento el espaciado indicado para cada pregunta. Si este es en formato digital, los márgenes, el interlineado, fuente y tamaño de letra vienen dados por defecto y no deben modificarse. En cualquier caso, asegúrate de que la presentación es suficientemente clara y legible. Entrega toda la documentación relativa al examen, revisando con detenimiento que los archivos o documentos son los correctos. El envío de archivos erróneos o un envío incompleto supondrá una calificación de "no presentado".
4. Durante el examen y en la corrección por parte del docente, se aplicará el Reglamento de Evaluación Académica de UNIR que regula las consecuencias derivadas de las posibles irregularidades y prácticas académicas incorrectas con relación al plagio y uso inadecuado de materiales y recursos.
5. No está permitido el uso de Internet ni ningún tipo de comunicación con otra persona. Durante todo el examen tu teléfono móvil debe estar en modo avión.
6. La parte principal de cada pregunta consiste en interpretar y comentar los resultados obtenidos. Si te limitas a hacer los cálculos no vas a poder superar el examen.
7. Es fundamental que las respuestas estén debidamente redactadas, de forma clara y precisa y sin faltas de ortografía.
8. Para hacer el examen puedes utilizar los apuntes del curso y los scripts que hayas preparado y Python para hacer los cálculos.

## Puntuación

### Preguntas

- Puntuación máxima 10.00 puntos

**NO UTILIZAR ESTA  
PARTE DE LA HOJA**


DATOS PERSONALES		FIRMA
Nombre: <i>Diego Alberto</i>	DNI: <i>0922789151</i>	
Apellidos: <i>Colame Borge</i>		

El examen constará de un ejercicio práctico (8,5 puntos) y una pregunta teórica (1,5 puntos). Los enunciados están en la página 14 y el espacio para responder el examen está entre las páginas 4 y 13.

#### 1. Pregunta

NO UTILIZAR ESTA  
PARTE DE LA HOJA



DATOS PERSONALES		FIRMA
Nombre: <u>Diego Alberto</u>	DNI: <u>6722787151</u>	
Apellidos: <u>Colome Borge</u>		

Pregunta 1.

fecha nacimiento: 29/04/1997

•  $m = 0.4 + 3 = 7$

$29 > 21 \Rightarrow d = 16$

Vamos a usar  $x_1, x_2, x_3, x_7, x_{16}$ , y

$x_1 = [-0.53, 0.80, 0.01, 1.39, 0.03, -0.46, -1.40, -0.74, -0.41, 1.85]$

$x_2 = [-1.69, -1.24, -1.01, -1.02, -0.01, 0.11, 0.65, 2.37, 0.35, -0.23]$

$x_3 = [0.26, 0.48, 0.61, -1.20, -1.23, -1.30, -1.14, 0.65, -1.34, 0.85]$

$x_7 = [-0.15, 0.37, 0.77, 1.64, -2.84, -0.09, -0.43, -0.88, 0.31, -1.64]$

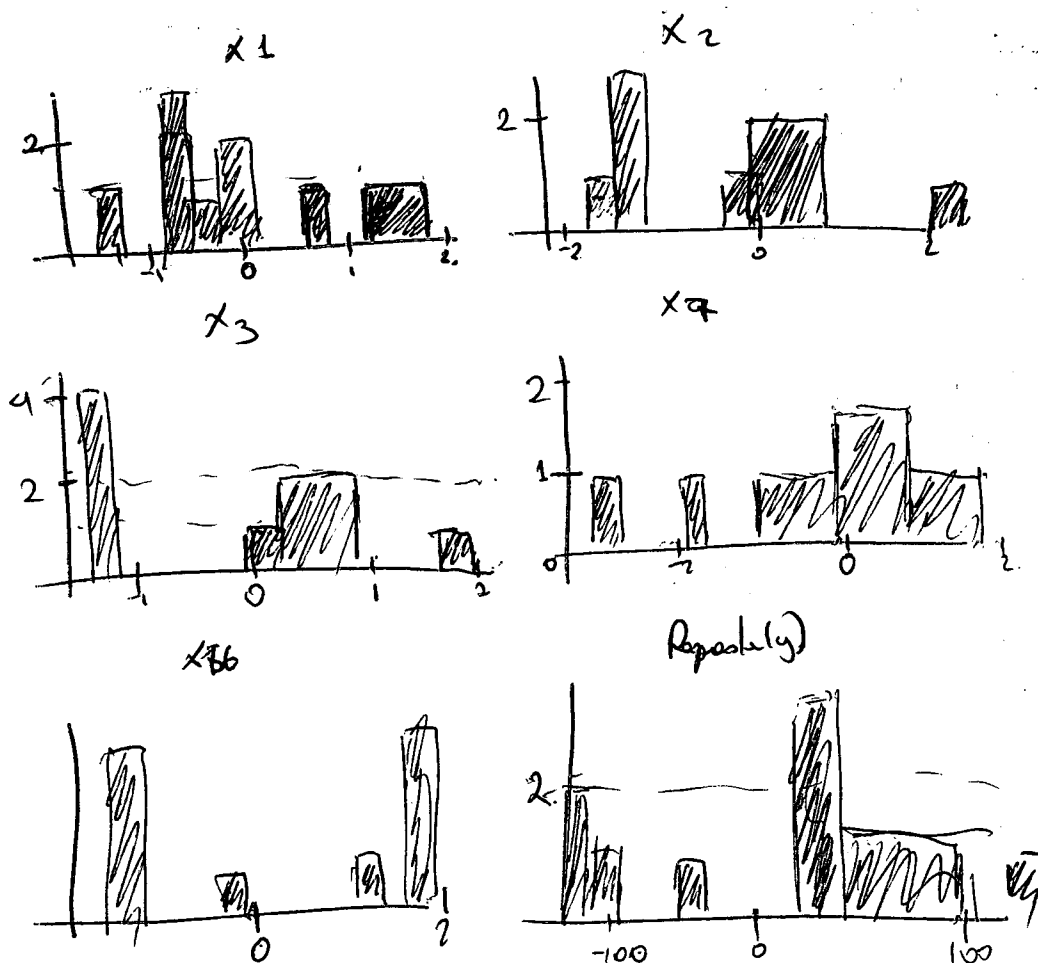
$x_{16} = [1.94, 1.72, 1.62, -0.18, 1.94, -1.47, -1.78, -1.72, -1.70, 1.08]$

$y = [-128.65, 24.66, 13.61, 46.52, 92.21, -32.29, -148.49, 11.51, -114.79, 171.13]$

→ Los datos se convierten a un dataframe y los resultados de `describe()` e `hist()` son:

	$x_1$	$x_2$	$x_3$	$x_7$	$x_{16}$	Respuesta ( $y$ )
mean	0.054	-0.17	-0.026	-0.289	0.14	-6.45
std	1.007	1.17	1.15	1.26	1.67	101.81
min	-1.40	-1.69	-1.36	-2.84	-1.78	-148.49
25%	-0.5125	-1.01	-1.18	-0.767	-1.64	-94.165
50%	-0.2	-0.12	0.37	-0.09	0.45	12.56
75%	0.6	0.79	0.64	0.35	1.64	41.05
max	1.85	2.37	1.93	1.64	1.94	171.13

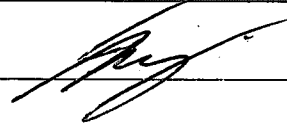
# histograms



Viendo los datos de las descripciones se evidencia que las variables  $X_1, X_2, X_3, X_7, X_{16}$  siguen distribuciones con medias cercanas al 0 y desviaciones cercanas al 1, por lo que se podría esperar encontrar distribuciones normales estándar.

Sin embargo, vamos a analizar la p-value de las distribuciones para cada variable descriptiva. Las variables  $X_1, X_2, X_3, X_7, X_{16}$  podrían considerarse como la más cercana a una distribución normal, mientras que la  $X_{16}$  sufre del caso contrario. Esto es interesante debido a que la Respuesta también parece seguir una distribución normal.

Esto nos indica que en efecto podrían existir datos anómalos en el dataset.

DATOS PERSONALES		FIRMA
Nombre: Diego Albano	DNI: 0922789151	
Apellidos: Colina Borja		

Y Separar el dataset para trabajar

training = df.iloc[:8]

test = df.iloc[8:]

2 Regresión Lineal multivariante.

Usando sklearn, el módulo Linear Regression se obtienen:

Se obtienen coeficientes:  $[98.05489526, 1268.391297, -1216.404, -13.18234234, 1239.038868]$   
intercepto:  $-0.5540737$

porque el modelo ajusta casi a perfección:

$$y = 98.05 X_1 + 1268.39 X_2 - 1216.404 X_3 - 13.18 X_4 + 1239.04 X_5 - 0.55$$

Además, utilizando statsmodels podemos fácilmente obtener los siguientes

resultados

$$R^2 = 0.899$$

$$R^2_{adj} = 0.64$$

	$P >  t $
X1	0.018
X2	0.799
X3	0.807
X4	0.657
X5	0.804

Vemos que hay un buen ajuste de los datos según el valor de  $R^2$ .

Aunque es muy bueno para

$R^2_{adj}$ . Además vemos que

el p-valor de X1 es muy significativo

mientras que el de los otros es muy alto

como podemos ver.

Es interesante dado que los coeficientes de

X2, X3 y X5 son muy altos. Quizás por un efecto de compensación.

## o Valores de VIF.

En la regresión anterior se obtuvieron los siguientes datos de VIF

Var	VIF
X1	1.864217.
X2	0.5350.0.18003
X3	2.1649.677595
X7	3.140585
X16	163959.019693

En esto se confirma algo que habíamos notado en los histogramas y es que  $X_1$  y  $X_7$  afectan la mayor cantidad de problemas. Mientras que  $X_{16}$  lo menor. Basados en esto se realiza el algoritmo de selección siguiente.

~~selección siguiente.~~

## 3 Selección StepWise

**1ero** vamos a eliminar la columna con el VIF más significativo del. del phone ( $X_{16}$ ), y volver a hacer el ajuste.

$\text{train} = \text{training\_drop}("X_{16}")$

Haciendo esto obtenemos los siguientes resultados.

$$R^2 = 0.895$$

$$R^2_{adj} = 0.765$$


Var	$P >  t $
X1	0.026
X2	0.126
X3	0.409
X7	0.549

Var	VIF
X1	1.851741
X2	1.455772
X3	2.306077
X7	3.058431

$$y = 98.82 X_1 + 31.65 X_2 + 19.96 X_3 + -14.19 X_7 - 0.059$$

Muchos Mejor! mientras que el  $R^2$  no cambio mucho, el  $R^2_{adj}$  si lo hizo. Igual que los p-valores de todas las variables, Aunque algunos siguen siendo buenos. Los coeficientes variaron mucho, pero no el de  $X_1$  que seguía el mejor ajuste antes. EN VIF de todas las predictoras por otra parte mejoró notablemente. Veamos que pasa con una segunda selección



DATOS PERSONALES		FIRMA
Nombre: <u>Diego Albaladejo</u>	DNI: <u>0922789157</u>	
Apellidos: <u>Colombes</u>		

2do. Vamos a eliminar ahora la variable  $X_3$  que tiene el mayor VIF

$\text{trany} = \text{lm}(\text{trany} \sim \text{drop}(columns = "X_3"))$

$$R^2 = 0.88$$

$$R^2_{adj} = 0.789$$

Var	$P >  t $
$X_1$	0.01
$X_2$	0.058
$X_3$	0.073

Var	VIF
$X_1$	1.3544
$X_2$	1.3511
$X_3$	1.0056

$$y = 91.14 X_1 + 34.88 X_2 + 30.51 X_3 + 2.28$$

Es claro que sucede algo similar al resto  $X_3$ . El  $R^2$  disminuye levemente, pero mejora  $R^2_{adj}$ . Los p-values son todos pequeños ahora de forma que podrían notarse como significativos y los VIFs se encuentran indudablemente todos en un rango.

Aunque sería posible continuar eliminando  $X_2$ , no lo consideramos necesario, a pesar de que se sigue un poco los VIF de  $X_1$  y  $X_2$ , no es destacable esto como para justificar la pérdida de información de este predictor. Aún así por completitud señalamos los resultados de

hacerlo.

3ro.  $\text{trany} = \text{lm}(\text{trany} \sim \text{drop}(columns = "X_2"))$

$$R^2 = 0.704$$

$$R^2_{adj} = 0.585$$

Var	$P >  t $
$X_1$	0.019
$X_2$	0.103

Var	VIF
$X_1$	1.34
$X_2$	1.30

$$y = 90.1425 X_1 + 37 X_2 + 9.1$$

Como habíamos mencionado, mejora el VIF, pero se pierde un buen pedacito de ajuste

Teniamos to en cuenta recomendamos para después de eliminar  $X_7$  como variable predictora. Aunque se podría hacer un argumento por para antes, consideramos que en este caso se obtiene el mejor balance.

## 4 Regresión LASSO

Utilizab. sklearn.linear-model.Lasso

Se obtiene el modelo:

coef: [95.4145 30.2555, 20.166 -12.9993 0]

inte: -15.11

$$\Rightarrow y = 95.41 X_1 + 30.25 X_2 + 20.16 X_3 - 12.999 X_4 - 15.11$$

con  $R^2 = 0.8943$

Es interesante notar que el modelo obtenido por machine Lasso ofrece un  $R^2$  bueno eliminando  $X_6$ . Es decir un  $R^2$  ligeramente mejor que cuando lo eliminamos manualmente.

5

## Comparativa

Para realizar la comparación, vamos a tener en cuenta los valores de los coeficientes, el  $R^2$ , el error medio el. predicc y la forma del modelo

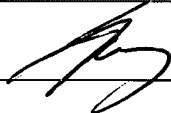
DATOS PERSONALES		FIRMA
Nombre:	Diego Alberto	
DNI:	0122789151	
Apellidos:	Cidani Borja	

Table comparison.

Regression lineal multiple.

Selección Stepwise

LASSO

Formulas	$y = b_1x_1 + b_2x_2 + b_3x_3 + b_7x_7 + b_{16}x_{16} + C$	$y = b_1x_1 + b_2x_2 + b_3x_3 + C$	$y = b_1x_1 + b_2x_2 + b_3x_3 + b_7x_7 + C$
Intercepto	-0.55	2.28	-15.11
$b_1$	98.05	95.14	95.91
$b_2$	1268.39	371.88	30.25
$b_3$	-1261.909	30.51	20.16
$b_7$	-13.18	0	-12.999
$b_{16}$	1239.05	0	0
$R^2$	0.899	0.880	0.894
MSE	40.78	38.19	46.42

Para comparar los modelos es bueno notar lo que hace bien cada uno. La Regresión lineal múltiple ofrece un resultado que notando por el  $R^2$  es el mejor resultado. Sin embargo, para lograrlo necesita usar coeficientes exagerados, y se observó un alto grado de influencia de la variable. En particular por parte del predictor  $x_{16}$  que es el más en los otros modelos.

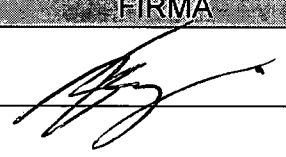
Siempre, dedique el intercepto <sup>es el</sup> ~~que~~ origen,  
algo que habíamos esperado desde el inicio. Podemos confirmar que  
el ajuste. Aunque sobreajustado, es bueno. El método de selección stepwise  
nos da un ajuste bueno, aunque es el más bogablos 3. Resonancia  
menor en merced variables, mostrando cuales realmente son los más  
significativos, y manteniendo el intercepto cerca del origen.

Por otra parte, el modelo obtenido por los ~~datos~~ encuentra un buen  
~~Pregunta: Datos~~ ~~entrenamiento~~

balance, aprovechando el predictor  $x_7$  de forma significativa  
mientras que mantiene el modelo simple al anterior pero con  
mejor ajuste.

Pregunta: Datos anómalos

Se pueden detectar es conveniente eliminar datos anómalos  
del modelo. Si no, es posible usar métodos robustos que  
resisten su influencia. Por otra parte, métodos de validación pueden  
servir para la detección.

DATOS PERSONALES		FIRMA
Nombre:	<i>Diego Alberto</i>	
DNI:	<i>0922789151</i>	
Apellidos:	<i>Colonna Borge</i>	

## Preguntas - Pregunta 1

### Ejercicio (8.5 puntos)

Escribe tu fecha de nacimiento (por ejemplo, 30/04/1987) y realiza los siguientes cálculos

- Calcula  $m$  como la suma de los dígitos del mes en que naciste y el número 3 (en el ejemplo anterior,  $m = 0 + 4 + 3 = 7$ )
- Calcula  $d$  en función del día que naciste. Si el día que naciste  $\in [1, 10]$ , entonces  $d = 14$ , si  $\in [11, 20]$  entonces  $d = 15$  y si por el contrario es  $\geq 21$  entonces  $d = 16$  (en el ejemplo anterior,  $30 \geq 21 \Rightarrow d = 16$ )

Para hacer el examen debes considerar un *dataset* formado por las variables  $x_1, x_2, x_3, x_m, x_d, y$ . (en el ejemplo anterior, serían las variables  $x_1, x_2, x_3, x_7, x_{16}, y$ ) de la Tabla ??

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	y
-0.53	-1.69	0.26	0.04	0.27	0.18	-0.15	-0.65	-0.52	-1.54	-2.91	1.37	0.25	-2.21	-0.02	1.94	-128.65
0.80	-1.24	0.48	0.38	0.55	-1.17	0.37	0.60	1.04	1.05	1.68	0.09	0.49	-0.45	1.76	1.72	24.66
0.01	-1.01	0.61	-0.81	-0.22	1.76	0.77	0.29	1.01	1.78	-1.32	0.51	-0.25	-1.00	1.23	1.62	13.61
1.39	-1.02	-1.20	-0.73	0.61	0.19	1.64	-0.10	0.54	1.80	0.23	0.53	-0.44	0.37	-1.01	-0.18	46.52
0.03	-0.01	1.93	-2.10	-0.82	-0.89	-2.84	0.04	-1.08	-1.16	-0.28	0.10	0.83	0.03	3.90	1.94	92.21
-0.46	0.11	-1.36	-0.43	-1.67	-0.47	-0.04	-1.83	-1.64	-0.22	-0.09	0.24	-0.25	-0.34	-3.17	-1.47	-32.29
-1.40	0.65	-1.14	-0.91	1.40	0.31	-0.43	0.24	0.30	-0.51	0.42	3.39	1.27	-0.76	-3.68	-1.78	-148.49
-0.74	2.37	0.65	-0.90	0.47	-0.88	-0.88	1.74	1.22	0.47	-1.93	-0.35	-0.49	1.63	0.56	-1.72	11.51
-0.41	0.35	-1.34	1.93	-1.60	-0.95	0.31	0.65	-0.21	1.54	-0.23	-0.33	0.28	-0.06	-3.10	-1.70	-114.79
1.85	-0.23	0.85	-0.13	1.27	0.37	-1.64	-0.24	-0.73	2.53	1.46	-0.40	0.67	1.61	3.54	1.08	171.13

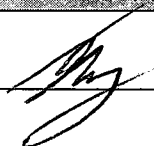
Tabla 1: conjunto general de variables

Contesta a las siguientes preguntas:

1. (2 puntos) Representar los datos: analizar si hay posibles datos anómalos, tablas y gráficas a modo de resumen que se consideren relevantes. Divide el *dataset* en entrenamiento (0.8) + test (0.2). Interpreta y comenta los resultados obtenidos
2. (1.5 puntos) Realiza un ajuste de regresión lineal multivariante para predecir la variable  $y$  a partir del resto y comenta los resultados obtenidos
3. (1.5 puntos) Realiza una selección de variables con el método *stepwise* atendiendo al valor del VIF y comenta los resultados obtenidos
4. (1.5 puntos) Realiza una regresión con el método de LASSO y comenta los resultados obtenidos
5. (2 puntos) Compara los resultados obtenidos en los 3 ajustes atendiendo a los valores de  $R^2$ ,  $R^2$  ajustado y el error cuadrático medio, además de a otros criterios a tu elección. Explica de forma razonada qué método crees que es mejor

### Pregunta (1.5 puntos)

Explica en menos de 200 palabras el problema que puede acarrear tener datos anómalos en la muestra. Debes contestar a las preguntas ¿qué consecuencias tiene?, ¿cómo podemos detectarlo?, ¿cómo podemos trabajar si tenemos una muestra con datos anómalos?

DATOS PERSONALES		FIRMA
Nombre:	<i>Diego Alberto</i>	
DNI: <i>0922789151</i>		
Apellidos:	<i>Colome Borge</i>	

**B O R R A D O R**  
**PÁGINA NO VÁLIDA PARA RESPONDER**

**B O R R A D O R**  
**PÁGINA NO VÁLIDA PARA RESPONDER**