| DATOS PERSONALES | FIRMA |
| --- | --- |
| |  |  | | --- | --- | | Nombre: Leonardo | DNI: 43470416W | |  |
| Apellidos: Prieto Yagi |

| 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 - B | ONLINE |

| Etiqueta identificativa |
| --- |
|  |

**INSTRUCCIONES GENERALES**

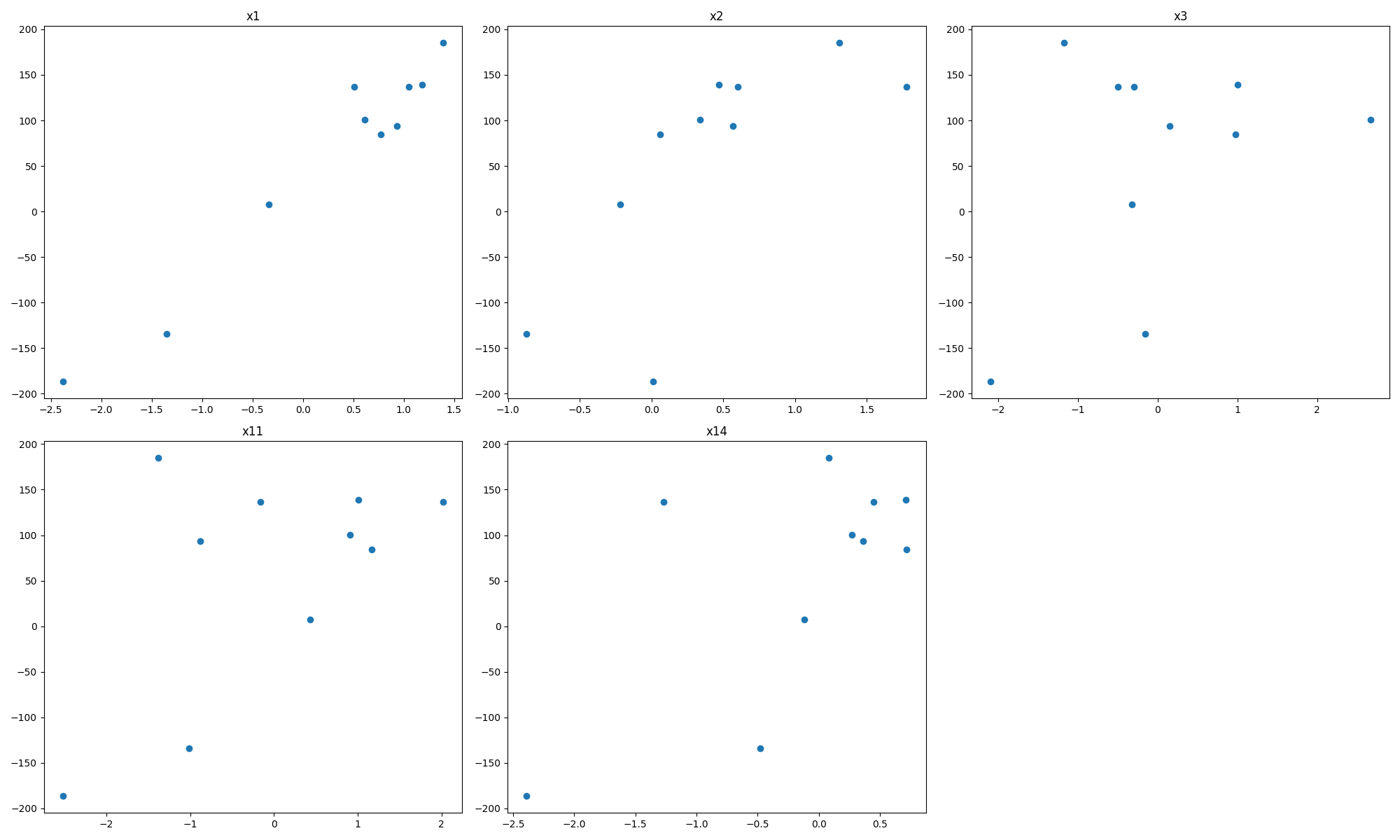
1. Ten disponible tu documentación oficial para identificarte, en el caso de que se te solicite.
2. Si tu examen consta de una parte tipo test, indica las respuestas en la plantilla según las características de este.
3. Debes contestar en el documento adjunto, respetando en todo momento el espaciado 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**

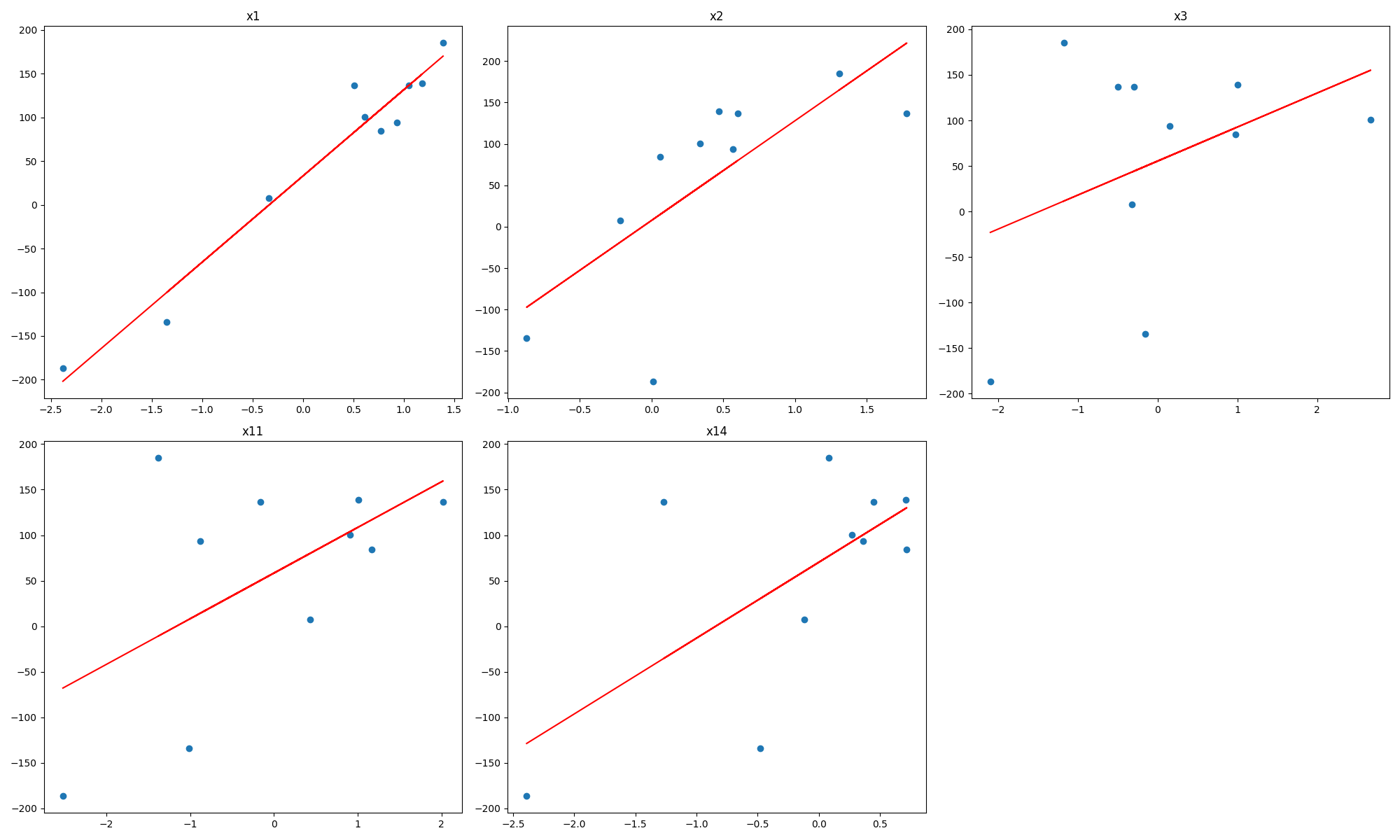
**Preguntas**

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 práginas 4 y 13.  
  
**1.** Pregunta

Mi fecha de nacimiento es 03/09/1996. Por tanto m=11 y d=14. Se toman entonces las variables predictoras x1, x2, x3, x11 y x14. Empezamos representando los datos:



Con regresión lineal en cada una de las variables se vería de la siguiente manera



 Dado que cada variable contiene muy pocos datos, mostrar un histograma no resulta muy relevante, disponiendo ya de las gráficas. Se puede ver cómo x1 y x2 presentan una relación lineal respecto a y, mientras que para x3, x11 y x14, esta relación no parece muy acertada.

En x2, donde la tendencia general de los datos es lineal, se puede observar un dato anómalo respecto a dicha tendencia alrededor de x2. En x11 y x14 uno se podría atrever a decir que hay una cierta tendencia lineal, exceptuando 3 o 4 datos anómalos. Sin embargo, tal cantidad de datos representa un porcentaje muy elevado del número de muestras, lo cual hace plantearse que asumir la relación lineal pueda no ser lo más correcto y que serían necesarios más datos para confirmarlo.

La descripción de los datos es la siguiente:

x1 x2 x3 x11 x14 y

count 10.000000 10.000000 10.000000 10.000000 10.000000 10.000000

mean 0.237000 0.405000 0.024000 -0.041000 -0.167000 56.296000

std 1.227364 0.750899 1.307093 1.396046 0.984593 123.845277

min -2.380000 -0.870000 -2.100000 -2.520000 -2.390000 -186.450000

25% -0.127500 0.022500 -0.455000 -0.977500 -0.390000 26.732500

50% 0.690000 0.405000 -0.230000 0.135000 0.175000 97.135000

75% 1.020000 0.592500 0.772500 0.985000 0.427500 136.615000

max 1.390000 1.780000 2.670000 2.020000 0.720000 184.970000

En general, las variables predictoras (no teniendo en cuenta la respuesta y), están centradas alrededor de 0 y con varianza unitaria. Es por ello que no se aplicará ninguna estandarización a los datos.

La primera regresión que se realiza es la lineal multivariante:

#### Regresión lineal múltiple ####

Intercept: 26.583708379120385

Coefs: [66.50563103 49.0181251 -2.01439438 8.93035146 17.48750593]

OLS Regression Results

========================================================================

Dep. Variable: y R-squared: 0.992

Model: OLS Adj. R-squared: 0.982

Se observa que en los coeficientes, las variables que más pesan son x1 y x2 (las que mostraban mayor tendencia lineal), seguido de x14 con mucha diferencia. Finalmente, x11 tiene un peso bastante menor al resto y x3 es la que menos se tiene en cuenta. Esto encaja bastante bien con lo analizado a partir de las gráficas. Los valores de son elevados, lo cual indica que se ha conseguido ajustar los datos de entrenamiento.

Imagen de la pantalla de un celular de un mensaje en letras blancas

Descripción generada automáticamente con confianza baja

Los p-valores muestran concordancia con los coeficientes de cada variable, siendo x3 y x11 los que muestran un valor mayor. Estos serían los candidatos a eliminar mediante el algoritmo. El ajuste obtenido tras llevar a cabo dicha eliminación es el siguiente:

Texto

Descripción generada automáticamente

Se obtiene un coeficiente simple mayor, pero el ajustado tiene el mismo valor. Los coeficientes de la regresión, de nuevo, dan mayor peso a x1 y x2. Como hemos visto, esto es porque dichas variables siguen siendo las que muestran una mayor tendencia lineal. Sin embargo, los p-valores son mucho menores respecto a los de la primera regresión, lo cual indica que hemos conseguido escoger un valor menor de variables que aportan la información suficiente. Dado que el es muy similar, este modelo sería preferible debido a su mayor simpleza.

Se pasa ahora a analizar el VIF, es decir, la colinealidad. Los resultados obtenidos son los siguientes:

Variable VIF

0 x1 156414.083182

1 x2 74396.848608

2 x3 2.616989

3 x11 2.062794

4 x14 99654.834807

Se aprecia que las variables x1, x2 y x14 muestran valores elevados. Por tanto, se deciden eliminar las variables x2 y x14

Texto

Descripción generada automáticamente

El resultado, atendiendo a los , es peor que los anteriores. La x1 sigue siendo la que mayor peso tiene en la regresión con diferencia. Esto es normal porque, como hemos visto, x3 y x11 no mostraban una tendencia lineal clara. Sin embargo, viendo que se sigue obteniendo un valor alto de , y teniendo en cuenta que de esta manera se pueden eliminar colinealidades entre variables, sería de interés estudiar el comportamiento del modelo con un mayor número de muestras. Si, finalmente, la variables que se usan acaban mostrando un mayor comportamiento lineal, este podría ser mejor que el modelo de eliminación sobre p-valores.

He tenido bastantes problemas técnicos y no tengo tiempo a hacer la parte de evaluación de RMSE… sin embargo, los modelos con mayor seguramente mostrarían menos error. La elección del mejor modelo la he intentado explicar en los apartados anteriores y no se vería muy alterada.

(Responder en 10 caras)

