| DATOS PERSONALES | FIRMA |
| --- | --- |
| |  |  | | --- | --- | | Nombre: Patricio Gonzalo | DNI: 1713941704 | |  |
| Apellidos: Bayas Meza |

| 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 | Quito, Ecuador |

| 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**

* Puntuación máxima 10.00 puntos

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

Fecha de nacimiento: 09/07/1977

m= 07+3=10

como el día de nacimiento es 09 entonces d=14

* + - 1. **Representar los datos**

El primero paso es realizar un análisis exploratorio de los datos, con este fin se aplican los métodos describe() e info() para conocer las características de los datos.

Presentación de los datos con los que se trabajará

Imagen de la pantalla de un computador

Descripción generada automáticamente con confianza baja

Revisados los datos con los proporcionados como base para el ejercicio se confirma que están correctamente importados desde el archivo .csv en el que se los copió inicialmente.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

El método describe () muestra que se tiene datos anómalos que se los debe estudiar esto se puede apreciar en vista que en todas las variables y en especial a la respuesta la desviación estándar alta con valores que se los puede confundir con los datos originales, por otra parte, la diferencia entre el valor máximo y el cuartil 3 (75%) es mayor a la diferencia entre los otros cuartiles.

Texto

Descripción generada automáticamente

El método info() muestra que todas las variables son del tipo flotante, no se encuentras datos que puedan alterar los análisis que se van a realizar.

Los datos tanto de las variables predictoras como respuesta se representan en histogramas:

Gráfico, Gráfico de barras, Histograma

Descripción generada automáticamente

Para interpretar de mejor manera el comportamiento de los datos se obtienen sus distribuciones.

En el histograma de X2 se observa que la frecuencia de los datos entre -1.5 y 1.0 es similar a la que presenta entre 0.00 y 0.05.

En todas las variables se presenta valles en los que no hay datos y ninguna gráfica muestra una similitud con una distribución normal, en parte por la cantidad de datos de cada variable.

En la variable x14 se encuentran frecuencias de datos alta en los extremos con lo que se puede inferir que los datos anómales se presentan en esta variable.

Variables predictoras (x1, x2, x3, x10, x14)

Gráfico, Histograma

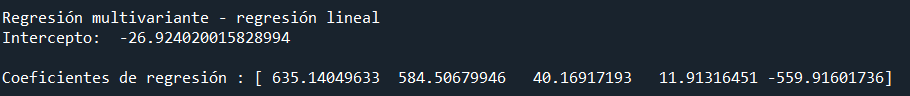
Descripción generada automáticamente

Las distribuciones muestran que los datos de las variables predictoras tienen una ligera cola hacia la derecha. Con el fin de apreciar de mejor manera el comportamiento de las variables se aplican transformaciones raíz cuadrada y logarítmica. Aplicadas las transformadas se aprecia que en la logarítmica se tienen una asimetría con cola hacia la izquierda.

Con las distribuciones se confirma que se cuentan con datos anómalos.

* + - 1. **Ajuste de regresión lineal multivariante**

Realizada la regresión lineal multivariante se tienen los siguientes resultados:



Se identifica un intercepto de alrededor de -26.92 junto con cinco coeficientes de regresión, entre los que se destaca un valor negativo cercano a -559.92.

Aplicado un resumen del modelo se encuentra:

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Se identifica que las variables predictoras no son estadísticamente significativas en vista que se tienen valores de P>|t| mayores a 0.05, por otra parte, en el intervalo [0.025; 0.975] en todos los casos el cero es parte del intervalo con lo que se confirma que las variables predictoras no son estadísticamente significativas.

Se puede observar que el coeficiente de determinación del modelo es igual a 0.973, y que el coeficiente de determinación ajustado es 0.905 con una diferencia de 0.068 que es provocada porque el número de observaciones es mayor que el número de variables del modelo.

Se puede apreciar que el error es similar en el caso de x1, x2, x14 y alrededor de 140 veces el error x3 y x10.

* + - 1. **Selección de variables con el método stepwise**
      2. **Regresión con el método de Lasso**

Aplicado el método de Laso se tiene como resultados:

Interfaz de usuario gráfica, Texto, Chat o mensaje de texto

Descripción generada automáticamente

Se identifica un intercepto de alrededor de -26.08 muy similar al encontrado en la regresión lineal, en lo que corresponde a los coeficientes de regresión, para las variables x2 es cero por lo que se puede afirmar que para el método de Laso esta variable no tiene participación, llama la atención que en el último coeficiente el valor es positivo mientras en la regresión lineal este valor era negativo.

Un comparativo entre la regresión lineal y el método de Lasso se presenta a continuación

Gráfico

Descripción generada automáticamente

Se aprecia que el método lineal tiene una mayor variación de beta que el modelo de Lasso

* + - 1. **Comparación**

Realizada la comparación se tienen los siguientes resultados:

Texto

Descripción generada automáticamente

Se puede apreciar que el método de Lasso tiene un menor error cuadrático medio

**2. Que puede acarrear tener datos anómalos en la muestra**

¿Qué consecuencias tiene?

La presencia de datos anómalos en las muestras no debe ser tratados como errores, se los debe identificar y analizar porque se presentan, lo que se debe cuidar es las herramientas que se apliquen para interpretar los datos o para representarlos, sin esta consideración lo que se estime acarreará un error que se aplicará conforme se apliquen transformaciones a los datos.

Por ejemplo, si los datos anómalos provocan una desviación estándar con un valor similar al de los datos es un error utilizar la media aritmética como una medida representativa de los datos, en este caso se debe considerar utilizar la mediana o la moda según lo que se desea analizar.

¿Cómo podemos detectarlo?

Se los puede detectar con herramientas estadísticas como histogramas, diagramas de caja y bigotes (box plot), análisis de intervalos, diagramas de distribuciones, etc., el correcto uso de las herramientas que brinda la estadística son importantes para identificar los datos anómalos.

¿Cómo podemos trabajar si tenemos una muestra con datos anómalos?

Se debe identificar y confirmar que es un dato anómalo, acercarnos al sistema que entregó el dato, analizar las condiciones en las que se generan los datos, establecer si se tiene una frecuencia del dato anómalo, y finalmente emplear las herramientas estadísticas que no se vena afectadas por datos anómalos o extremos como medianas, ajustes a regresiones, aplicación de métodos como stepwise, etc.

