Estadística II - MGAIE

Ejercicios de Regresión lineal

EJERCICIO 1. ESTADISTICA DESCRIPTIVA Y CORRELACION

Por medio de este ejercicio se propone evaluar la capacidad de análisis de asociación y de predicción entre variables cuantitativas. Para este fin se utilizará el conjunto de datos que se describe a continuación: Archivo **CORDOBA.SAV**

El archivo CORDOBA.SAV contiene información de algunas características de vivienda, hogar y población de departamentos de la provincia de Córdoba.

Este archivo consta de las siguientes variables:

- **DEPART**: Departamento provincial
- **NBI1**: Porcentaje de hogares con Necesidades Básicas Insatisfechas sobre el total de hogares de cada departamento.
- **NBI2**: Porcentaje de población en hogares con Necesidades Básicas Insatisfechas sobre el total de población de cada departamento.
- **CALMAT1**: porcentaje de viviendas que presentan materiales resistentes y sólidos en **todos** los paramentos (pisos, paredes o techos) e incorpora **todos** los elementos de aislación y terminación.
- **CALMAT2**: porcentaje de viviendas que presentan materiales resistentes y sólidos en **todos** los paramentos pero le faltan elementos de aislación o terminación **al menos en uno** de sus componentes (pisos, paredes, techos).
- **CALMAT3**: porcentaje de viviendas que presentan materiales resistentes y sólidos en **todos** los paramentos pero le faltan elementos de aislación o terminación en **todos** sus componentes, o bien presenta techos de chapa de metal o fibrocemento.
- **CALMAT4**: porcentaje de viviendas que presentan materiales no resistentes ni sólidos o de desecho **al menos en uno** de los paramentos.
- **ALFAB** : Porcentaje de alfabetos
- **COB** : Porcentaje de personas que tienen Obra social y/o plan de salud privado o mutual

Nota:

Las Necesidades Básicas Insatisfechas fueron definidas según la metodología utilizada en "La pobreza en la Argentina"

(Serie Estudios INDEC. N° 1, Buenos Aires, 1984).

Los hogares con Necesidades Básicas Insatisfechas (NBI) son los hogares que presentan al menos uno de los siguientes

indicadores de privación:

1- Hacinamiento: hogares que tuvieran más de tres personas por cuarto.

2- Vivienda: hogares en una vivienda de tipo inconveniente (pieza de inquilinato, vivienda precaria u otro tipo, lo que excluye

casa, departamento y rancho).

3- Condiciones sanitarias: hogares que no tuvieran ningún tipo de retrete.

4- Asistencia escolar: hogares que tuvieran algún niño en edad escolar (6 a 12 años) que no asistiera a la escuela.

5- Capacidad de subsistencia: hogares que tuvieran cuatro o más personas por miembro ocupado y, además, cuyo jefe no haya completado tercer grado de escolaridad primaria.

Se requiere realizar los siguientes análisis estadísticos sobre este conjunto de variables:

1. Realizar un análisis exploratorio de todas las variables de interés.
    a. Mediante los gráficos de Boxplot e Histograma, evaluar simetría, presencia de observaciones atípicas, dispersión de cada variable.
    b. Calcular indicadores de tendencia central y de dispersión, indicando cuál/cuáles son los más adecuados a las características de las variables.
2. Realizar gráficos de dispersión de la variable NBI con cada una de las restantes variables.
3. Calcular correlaciones lineales de la variable NBI con cada una de las variables disponibles.
4. Calcular la recta de regresión utilizando como variable dependiente el NBI y como independiente cada una de las restantes variables

EJERCICIO 2. PARA EJERCITAR EL CUERPO

Una medida objetiva del ajuste aeróbico de una persona es el consumo de oxígeno en volumen por peso unitario del cuerpo, por unidad de tiempo. Se utilizaron 31 individuos en un experimento con el objeto poder modelar el consumo de oxígeno mediante las siguientes variables:

- X1: Edad en años
- X2: Peso en kg.
- X3: Tiempo en recorrer 3 km
- X4: pulso en reposo
- X5: pulso al final del ejercicio
- X6: pulso máximo durante el ejercicio

Utilizando los datos del archivo “un ejemplo de regresion.sav”, resuelva las siguientes consignas:

1. Cuál es el nivel de asociación lineal de las variables con el consumo de oxígeno. Evalúe numérica y gráficamente. Comentar.
2. Realizar una regresión lineal con la variable X5.
3. Qué información da el coeficiente de determinación?
4. El parámetro de la variable independiente es significativo? Cuáles son los supuestos necesarios para definir esta prueba inferencial?
5. Realizar una regresión lineal múltiple, seleccionando los mejores predictores entre las variables independientes disponibles, utilizando un método de selección automática. Describir el proceso de selección automática utilizado. (Sug. Considerar como probabilidad de entrada 0.10 y de salida del modelo 0.15.
6. Analizar sobre la bondad del ajuste del modelo obtenido, comentando los indicadores y/o test que considera.
7. Realizar un análisis de los residuos del modelo para evaluar el cumplimiento de los supuestos. Para esto, realizar gráficos de los residuos con el valor predicho.
8. Analizar la colinealidad de las variables predictoras presentes en la ecuación.
9. Analizar la presencia de observaciones atípicas y/o influyentes. Comentar y resolver según el caso.

EJERCICIO 3. Tasa de homicidios en Detroit

En un estudio que investiga el papel de las armas de fuego en el aumento de la tasa de homicidios de Detroit, se recogieron datos para los años 1961-1973. La variable de respuesta (la tasa de homicidios) y las variables potencialmente predictoras del comportamiento de ésta, se describen a continuación (archivo p301.sav):

Variable	Descripción
FTP	Cantidad de policias full time cada 100.000 habitantes
UEMP	Porcentaje de desempleados
M	Cantidad de trabajadores en la Industria Manufacturera (en miles)
LIC	Cantidad de licencias de portación de armas cada 100.000 habitantes
GR	Cantidad de armas de fuego registradas cada 100.000 habitantes
CLEAR	Porcentaje de homicidios resueltos con arresto del responsable
W	Cantidad de hombres blancos en la población
NMAN	Cantidad de trabajadores fuera de la industria manufacturera
G	Cantidad de trabajadores del gobierno
HE	Ingreso promedio por hora
WE	Ingreso promedio semanal
H	Tasa de homicidios cada 100.000 habitantes

Se requiere construir una ecuación de regresión que relacione la Tasa de Homicidios (variable H), con el resto de las variables disponibles, y determinar si estas variables son útiles para predecir la Tasa de Homicidios.

1. Cuál es el nivel de asociación lineal de las variables predictoras con la variable H? Comentar.
2. Realizar una regresión lineal múltiple, seleccionando los mejores predictores entre las variables independientes disponibles, utilizando un método de selección automática. Describir el proceso de selección automática utilizado. (Sug. Considerar como probabilidad de entrada 0.10 y de salida del modelo 0.15.)
3. Qué información da el coeficiente de determinación?
4. Cuáles son los supuestos necesarios para definir la prueba inferencial de los estimadores de los parámeros?
5. Analizar la bondad del ajuste del modelo obtenido, comentando los indicadores y/o test que considera.
6. Realizar un análisis de los residuos del modelo para evaluar el cumplimiento de los supuestos. Para esto, realizar gráficos de los residuos con el valor predicho.
7. Analizar la colinealidad de las variables predictoras presentes en la ecuación.
8. Analizar la presencia de observaciones atípicas y/o influyentes. Comentar y resolver según el caso.

EJERCICIO 4. Consumo de cigarrillos

El siguiente conjunto de datos refieren al consumo de cigarrillos en los Estados Unidos. La base de datos que se presenta (archivo p081.sav) corresponde a los datos de los 50 estados del país, y las variables consignadas son las siguientes:

Variable	Definición
Age	Edad mediana de las personas que viven en el Estado
HS	Porcentaje de personas de más de 25 años con secundario completo
Income	Ingreso personal per capital
Black	Porcentaje de indivduos de raza negra
Female	Porcentaje de mujeres
Price	Precio promedio de un paquete de cigarrillos
Sales	Número de paquetes vendidos per cápita en el estado

Se requiere construir una ecuación de regresión que relacione el consumo de cigarrillos per-cápita en todo el estado (variable Sales), con diversas variables socioeconómicas y demográficas, y determinar si estas variables son útiles para predecir el consumo de los cigarrillos.

1. Cuál es el nivel de asociación lineal de las variables predictoras con la variable Sales? Comentar.
2. Realizar una regresión lineal múltiple, seleccionando los mejores predictores entre las variables independientes disponibles, utilizando un método de selección automática. Describir el proceso de selección automática utilizado. (Sug. Considerar como probabilidad de entrada 0.10 y de salida del modelo 0.15.)
3. Qué información da el coeficiente de determinación?
4. Cuáles son los supuestos necesarios para definir la prueba inferencial de los estimadores de los paráameros?
5. Analizar la bondad del ajuste del modelo obtenido, comentando los indicadores y/o test que considera.
6. Realizar un análisis de los residuos del modelo para evaluar el cumplimiento de los supuestos. Para esto, realizar gráficos de los residuos con el valor predicho.
7. Analizar la colinealidad de las variables predictoras presentes en la ecuación.
8. Analizar la presencia de observaciones atípicas y/o influyentes. Comentar y resolver según el caso.

EJERCICIO 5. Dimensiones del cuerpo

El siguiente conjunto de datos refieren a una serie de medidas corporales y esqueléticas, edad, peso, altura y sexo de una muestra de 507 personas físicamente activas. La base de datos que se presenta (archivo body.sav) contiene las siguientes variables:

Medidas del esqueleto

- V1: Diámetro Biacromial
- V2: Biiliac diameter (Amplitud pélvica)(ver figura)
- V3: Bitrochanteric diameter (ver figura)
- V4: Profundidad del pecho entre la columna vertebral y el esternón a nivel del pezón
- V5: Diámetro del pecho a nivel del pezón
- V6: Diámetro del codo, suma de los dos codos
- V7: Diámetro de las muñecas, suma de amabs
- V8: Diámetro de las rodillas, suma de ambas
- V9: Diámetro del tobillo, suma de ambos

Medidas del tipo circunferencias

- V10: Circunferencia de los hombros a la altura del músculo deltoides
- V11: Circunferencia del pecho a la altura de los pezones para los hombres, justo por encima del inicio del busto, e las mujeres
- V12: Circunferencia de la cintura justo debajo de la caja torácica
- V13: Circunferencia abdominal a la altura del ombligo
- V14: Circunferencia de la cadera
- V15: Circunferencia del muslo debajo del pliegue del glúteo. Pormedio de ambas piernas
- V16: Circunferencia del bicep flexionado. Promedio de ambos brazos
- V17: Circunferencia del antebrazo extendido. Promedio de ambos brazos
- V18: Circunferencia de la rodilla a la altura de la rótula con la pierna ligeramente flexionada. Promedio de ambas rodillas
- V19: Calf maximum girth, average of right and left girths
- V20: Circunferencia del tobillo. Promedio de ambos tobillos
- V21: Circunferencia de la muñeca. Promedio de ambas

Otras variables

- V22: Eda den años
- V23: Peso kg)
- V24: Altura(cm)
- V25: Sexo (1 - varón, 0 - mujer)

Se requiere construir una ecuación de regresión que relacione el peso con el resto de las variables, y determinar si estas variables son útiles para predecir el peso de los individuos.

1. Cuál es el nivel de asociación lineal de las variables predictoras con la variable Peso? Comentar.

2. Realizar una regresión lineal múltiple, seleccionando los mejores predictores entre las variables independientes disponibles, utilizando un método de selección automática. Describir el proceso de selección automática utilizado. (Sug. Considerar como probabilidad de entrada 0.10 y de salida del modelo 0.15.)
3. Qué información da el coeficiente de determinación?
4. Cuáles son los supuestos necesarios para definir la prueba inferencial de los estimadores de los paráameros?
5. Analizar la bondad del ajuste del modelo obtenido, comentando los indicadores y/o test que considera.
6. Realizar un análisis de los residuos del modelo para evaluar el cumplimiento de los supuestos. Para esto, realizar gráficos de los residuos con el valor predicho.
7. Analizar la colinealidad de las variables predictoras presentes en la ecuación.
8. Analizar la presencia de observaciones atípicas y/o influyentes. Comentar y resolver según el caso.