

**FACULTAD DE INGENIERÍA Y ARQUITECTURA**

**CARRERA DE INGENIERÍA DE SISTEMAS**

**ESTADÍSTICA APLICADA - SECCIÓN 203**

**2022 – 2**

**ESTUDIO DE CASO Nº 2**

**Caso: Consumo de agua en Lima**

Según el reporte de consumo doméstico de agua potable de cada habitante por día, los distritos de San Isidro, La Molina, Miraflores, Santiago de Surco, y San Borja registran consumos elevados que llegan hasta los 243 litros por persona, mientras que distritos como Carabayllo, Pachacámac, Ventanilla y Lurigancho (Chosica) no superan los 120 litros por persona, que es el indicador anhelado.

Al haber estas diferencias, Sedapal desea realizar un estudio para predecir el consumo de agua de la población y así crear una campaña para sensibilizar y concientizar a la población de Lima y Callao en el uso racional del agua potable, a no desperdiciarla y a generar hábitos de consumo responsable para darle sostenibilidad al servicio.

Para realizar el estudio Sedapal tomó una muestra aleatoria de viviendas al azar de 2 distritos y evaluó los consumos mensuales de agua de las familias y recogió información para las siguientes variables:

**Y =** Consumo mensual de agua (metros cúbicos)

**X1 =** Área de la vivienda (metros cuadrados)

**X2 =** Número de personas que habitan la vivienda

**X3 =** Número de habitaciones de la vivienda

X4= Antigüedad de la casa

X5= Distrito

X6= Tipo de material

Los datos se encuentran en el archivo ***“Consumo\_Agua.xlsx”***

1. Estimar la recta de regresión lineal que puede usarse para predecir el consumo mensual de agua dado el área de la vivienda. **(4 puntos)**
   1. Plantear el modelo e interpretar los parámetros estimados. **(2 punto)**

**Y = B0 + B1\*X1**

**Consumo = 6.80984 + 0.09068\*Área**

**B1 = 0.09068: Por cada metro cuadrado el consumo de agua aumenta en 0.09068 metros cúbicos.**

* 1. ¿Qué porcentaje de la variación total en el consumo mensual de agua se puede explicar mediante el modelo de regresión? **(1 puntos)**

**Multiple R-squared: 0.5313**

**El 53.13% de la variabilidad del consumo de agua es explicado por el área de la vivienda a través del modelo de regresión.**

* 1. Si una familia tiene una casa cuya área equivale a 100 m², ¿cuánto se esperaría que fuera su consumo mensual de agua? **(1 puntos)**

**15.87749 metros cúbicos de agua**

**Esta estimación no es tan confiable ya que el consumo solo depende en el 53% (de 90 a más es mejor).**

1. Realizar un modelo de regresión múltiple, tomando en cuenta lo siguiente:
   1. **Parte 1: Planteamiento del modelo inicial**  **(9 puntos)**
   2. Plantear el modelo de regresión lineal múltiple para estimar el consumo mensual en función de las variables Área de la vivienda, Número de personas que habitan la vivienda y Número de habitaciones de la vivienda. Estimar los parámetros del modelo planteado en a) e interpretar sus valores. **(4 puntos)**

Y = 9.180019 + 0.035244\*X1 + 0.422639\*X2 + 0.103001\*X3 + 0.004719\*X4

-1.097461\*X5Santa Anita - 0.623338\*X5Surco + 1.766375\*X6Otros

Si la casa es de Santa Anita se pone 1 ahí y 0 en Surco.

Si la casa es de Surco se pone 1 ahí y 0 en Santa Anita.

Si la casa es de otro que no Santa Anita ni Surco, se pone 0 en ambos.

Misma lógica para la variable X6Otros.

B2 = 0.422639: Por cada persona adicional que vive en la vivienda, el consumo de agua aumenta en 0.422639 metros cúbicos, manteniendo constante las otras variables.

B5 = -1.097461: Si la familia vive en San Anita, su consumo de agua disminuye en 1.097461 metros cúbicos, manteniendo constante las otras variables.

* 1. Evaluar si el modelo lineal múltiple es adecuado para predecir el consumo de agua mensual de una familia. Use **(2 puntos)**

H0: B1 = B2 = B3 = B4 = B5 = B6 = B7 = 0

H1: Al menos un Bi es diferente de 0

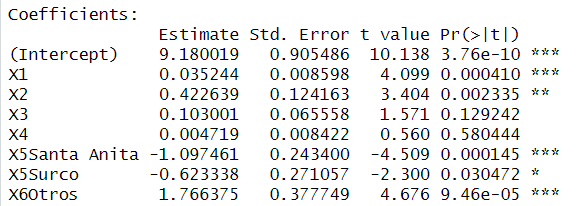
Se necesita el estadístico F

F-statistic: 59.12 on 7 and 24 DF, p-value: 1.361e-13

Como p-value < alfa, entonces se rechaza H0. Por lo tanto, al menos un Bi es diferente a 0, a un nivel de significancia del 0.05 (el modelo es válido y adecuado para predecir el consumo de agua mensual de una familia).

Adjusted R-squared: 0.9292. La influencia es mayor a 90%.

* 1. Determinar qué variables predictoras son significativas en el modelo lineal múltiple. Use **(3 puntos)**



Todas las variables son significativas menos X3 y X4 (por los \*). Por lo tanto, se quitan del modelo.

Esto es para el proyecto:

|  |  |  |  |
| --- | --- | --- | --- |
| H0: B1 = 0  H1: B1 ≠ 0  T = 4.099,  Pval = 0.000410  Rechaza H0 | H0: B2 = 0  H1: B2 ≠ 0  T = 3.404,  Pval = 0.002335  Rechaza H0 | H0: B3 = 0  H1: B3 ≠ 0  T = 1.571,  Pval = 0.1292  No se rechaza H0 | H0: B4 = 0  H1: B4 ≠ 0  T = 0.560,  Pval = 0.58044  No se rechaza H0 |
| H0: B5 = 0  H1: B5 ≠ 0  T = -4.509,  Pval = 0.000145  Rechaza H0 | H0: B6 = 0  H1: B6 ≠ 0  T = -2.3,  Pval = 0.030472  Rechaza H0 | H0: B7 = 0  H1: B7 ≠ 0  T = 4.676,  Pval = 9.46e-05  Rechaza H0 |  |

* 1. **Parte 2: Planteamiento del modelo final**  **(8 puntos)**

1. Estimar el modelo sólo con las variables predictoras significativas y responda lo siguiente:
   * 1. Calcular e interpretar el coeficiente de determinación múltiple adecuado.

**(1 puntos)**

Texto

Descripción generada automáticamente

Coeficiente de determinación: Adjusted R-squared: 0.9274

El 92.74% de la variabilidad total del consumo de agua es explicado por las variables en estudio a través del modelo de regresión.

* + 1. Estimar el consumo mensual de una familia cuando su casa tiene 100 m², en la vivienda habitan 5 personas y se cuenta con 4 cuartos, 8 años de antigüedad, es del distrito de “Surco” y de material “Noble” (considere solo las variables significativas). **(3 puntos)**

predict(modelo3, data.frame(X1 = 100, X2 = 5, X5 = "Surco", X6 = "Noble"))

# 14.84067

* + 1. Pruebe los supuestos de no multicolinealidad y la normalidad de los errores en el modelo. Use (Plantee las hipótesis y conclusiones respectivas). **(3 puntos)**
       - **Supuesto de no multicolinealidad**

Gráfico

Descripción generada automáticamente con confianza baja

Como los VIF < 4, no existe multicolinealidad.

* + - * **Supuesto de normalidad de los errores**

H0: Los errores se distribuyen normalmente

H1: Los errores no se distribuyen normalmente

Texto

Descripción generada automáticamente

Como p-value > alfa, no se rechaza H0. Entonces los errores se distribuyen normalmente, a un nivel de significancia del 0.05.

* + - * **Supuesto de Homogeneidad de varianza**
      * **Supuesto de no autocorrelación**

NOTA:

* Subir el archivo en Word
* Subir los códigos en R