Proyecto Regresión Lineal

Ariana Marcela Andrade Bello

Emanuel Esteban Restrepo Patarroyo

Universidad Unicomfacauca

Ingeniería de Sistemas

Estadística y Probabilidad

Popayán, Cauca

2024

Contenido

Introducción

La regresión lineal es una técnica fundamental en estadística y análisis de datos utilizada para modelar la relación entre dos o más variables. Este proyecto tiene como objetivo aplicar los principios de la regresión lineal para predecir y analizar datos relacionados con el consumo calórico diario y el peso corporal. Se analizará la relación entre estas variables mediante la aplicación de modelos estadísticos, interpretando los resultados y evaluando el modelo ajustado.

Objetivos

1. Recoger y preparar un conjunto de datos.
2. Aplicar un modelo de regresión lineal para entender la relación entre el consumo calórico diario y el peso corporal.
3. Evaluar el desempeño del modelo utilizando métricas estadísticas.
4. Interpretar los resultados obtenidos para hacer predicciones y recomendaciones basadas en el análisis.

Definición del Problema

El problema de investigación consiste en analizar cómo el consumo calórico diario (variable independiente, `X`) influye en el peso corporal (variable dependiente, `Y`) de una muestra de personas. Se busca determinar si existe una relación lineal significativa entre estas variables.

Recolección de Datos

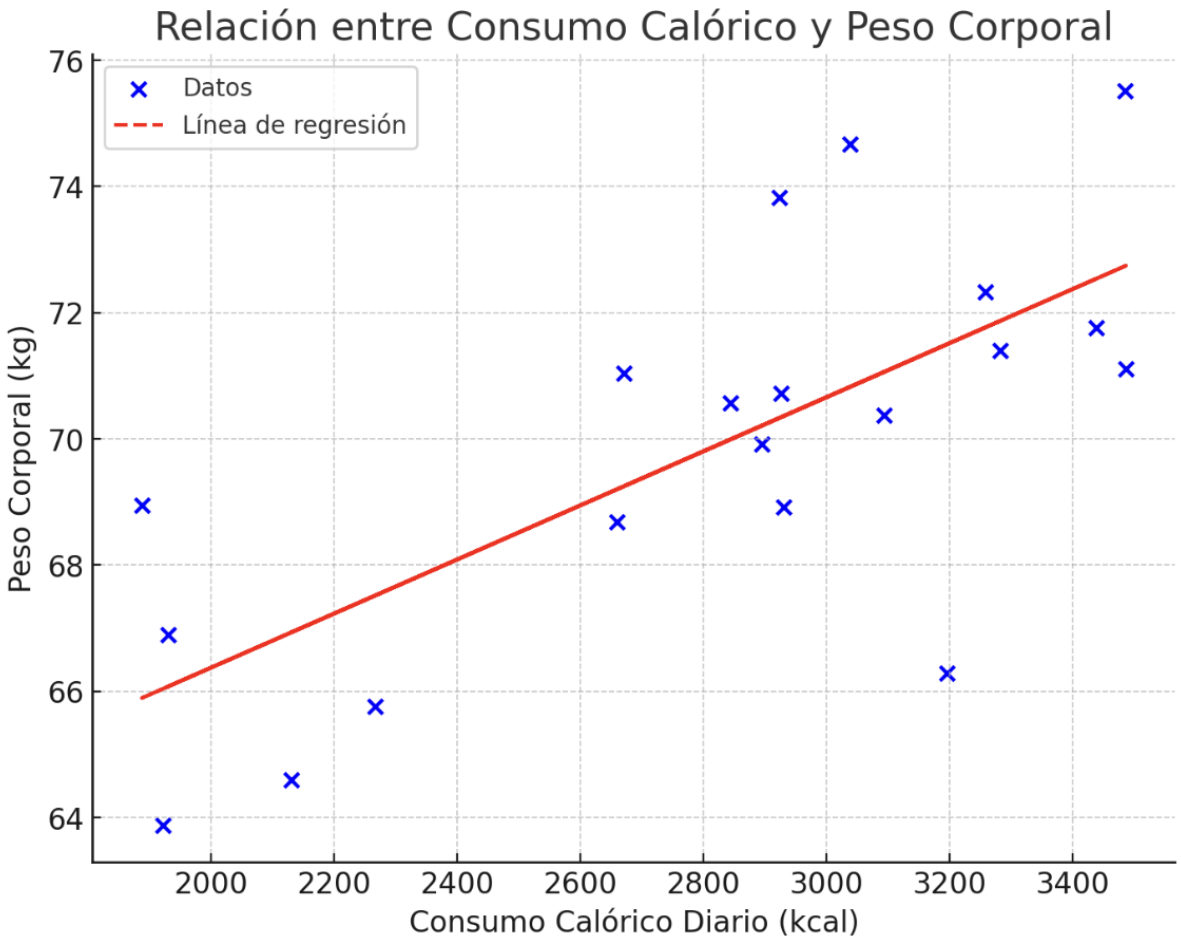
Se dispone de los siguientes datos obtenidos de 20 participantes:

Consumo Calórico (X)	Peso Corporal (Y)
2926	70.72
3259	72.33
2660	68.69
3094	70.37
2930	68.91
2895	69.91
2844	70.57
3438	71.76
1921	63.87
2266	65.76
3038	74.67
2130	64.60
3282	71.40
1887	68.95
3196	66.29
2923	73.82
2671	71.03
3487	71.11
1930	66.90
3485	75.51

Análisis Exploratorio de los Datos (EDA)

Visualización de los Datos

Se generó un gráfico de dispersión para visualizar la relación entre el consumo calórico diario y el peso corporal. El patrón general sugiere una relación lineal positiva.



Estadísticas Descriptivas

Variable	Media	Mediana	Desviación Estándar	Mínimo	Máximo	Rango
Consumo Calórico (X)	2813.10	2924.50		1887	3487	1600
Peso Corporal (Y)	69.86	70.47	3.18	63.87	75.51	11.64

Peso Corporal (Y)	Error respecto a la media	Error al cuadrado	Error cuadrado por frecuencia
70.72	0.8615	0.7422	0.7422
72.33	2.4715	6.1083	6.1083
68.69	-1.1685	1.3654	1.3654
70.37	0.5115	0.2616	0.2616
68.91	-0.9485	0.8997	0.8997
69.91	0.0515	0.0027	0.0027
70.57	0.7115	0.5062	0.5062
71.76	1.9015	3.6157	3.6157
63.87	-5.9885	35.8621	35.8621
65.76	-4.0985	16.7977	16.7977
74.67	4.8115	23.1505	23.1505
64.60	-5.2585	27.6518	27.6518
71.40	1.5415	2.3762	2.3762
68.95	-0.9085	0.8254	0.8254
66.29	-3.5685	12.7342	12.7342
73.82	3.9615	15.6935	15.6935
71.03	1.1715	1.3724	1.3724
71.11	1.2515	1.5663	1.5663
66.90	-2.9585	8.7527	8.7527
75.51	5.6515	31.9395	31.9395

Varianza

$$s^2 = \frac{192.2764}{19} \approx 10.1209$$

Desviación Estándar

$$\sqrt{s^2} = \sqrt{10.1209} \approx 3.1805$$

Identificación de Valores Atípicos

No se detectaron valores extremos que requieran eliminación o corrección. Todos los datos están dentro de un rango razonable.

Preparación de los Datos

- **Limpieza de datos**: No hay valores faltantes ni errores en el conjunto de datos.
- **División del conjunto de datos**: Se dividió el conjunto de datos en:
- **Conjunto de entrenamiento** (70%): 14 registros.
- **Conjunto de prueba** (30%): 6 registros.

Aplicación de la Regresión Lineal

Consumo Calórico (X)	Peso Corporal (Y)	X -Y	X^2	Y^2
2926	70,72	206926,72	8561476	5001,32
3259	72,33	235723,47	10621081	5231,63
2660	68,69	182715,40	7075600	4718,32
3094	70,37	217724,78	9572836	4951,94
2930	68,91	201906,30	8584900	4748,59
2895	69,91	202389,45	8381025	4887,41
2844	70,57	200701,08	8088336	4980,12
3438	71,76	246710,88	11819844	5149,50
1921	63,87	122694,27	3690241	4079,38
2266	65,76	149012,16	5134756	4324,38
3038	74,67	226847,46	9229444	5575,61
2130	64,60	137598,00	4536900	4173,16
3282	71,40	234334,80	10771524	5097,96
1887	68,95	130108,65	3560769	4754,10
3196	66,29	211862,84	10214416	4394,36
2923	73,82	215775,86	8543929	5449,39
2671	71,03	189721,13	7134241	5045,26
3487	71,11	247960,57	12159169	5056,63
1930	66,90	129117,00	3724900	4475,61
3485	75,51	263152,35	12145225	5701,76
56262,00	**1397,17**	**3952983,17**	**163550612**	**97796,42**

Link Código Fuente

Herramienta utilizada Python:

https://github.com/EstebanR05/regresion_lineal.git

Evaluación del Modelo

- **Ecuación del modelo de regresión lineal:**

$$\boldsymbol{Y = 57.815 + 0.00428X}$$

Donde:

$\boldsymbol{a = 57.815}$ es el intercepto.

$\boldsymbol{b = 0.00428}$ es el coeficiente de regresión.

- **Coeficiente de determinación ($\boldsymbol{R^2}$):**

$$\boldsymbol{R^2 = 0.5034}$$

Esto indica que el modelo explica aproximadamente el 50.34% de la variabilidad de Y (Peso Corporal) en función de X (Consumo Calórico).

- **Error Cuadrático Medio (RMSE):**

$$\boldsymbol{RMSE = 2.1846}$$

Esto mide la desviación promedio entre los valores observados y predichos en la misma unidad de **Y**.

- **Error Absoluto Medio (MAE):**

$$\boldsymbol{MAE = 1.7293}$$

Representa el error promedio absoluto entre las predicciones del modelo y los valores reales

Interpretación de los Resultados

Interpretación de los coeficientes:

- **Intercepto (**a=57.815**)**: Este valor representa el peso corporal promedio estimado (Y) cuando el consumo calórico (X) es igual a 0. Aunque no es un escenario realista (un consumo calórico de 0 no ocurre), es una referencia del punto inicial del modelo.
- **Coeficiente de regresión (**b=0.00428**)**: Esto indica que, por cada incremento de una unidad en el consumo calórico (X), el peso corporal (Y) se incrementa en **0.00428 kg**, en promedio. Esto sugiere que el consumo calórico tiene una relación positiva y directa con el peso corporal.

Evaluación del rendimiento del modelo:

- **Coeficiente de determinación** (**$R^2 = 0.5034$**)**:** Aproximadamente el 50.34% de la variación en el peso corporal (Y) está explicada por el consumo calórico (X). Aunque este valor muestra una relación moderada, hay un 49.66% de variación que no es explicada por este modelo, lo que sugiere que otras variables (como actividad física, metabolismo, etc.) pueden estar influyendo en el peso corporal.

Errores de predicción:

- **RMSE = 2.1846**: La desviación promedio entre los valores observados y los valores predichos es de 2.18 kg.

- **MAE = 1.7293:** En promedio, las predicciones del modelo se desvían de los valores reales en 1.73 kg.

Conclusión del modelo: Aunque el modelo muestra una relación significativa entre el consumo calórico y el peso corporal, no es completamente preciso. Se pueden incluir más variables independientes para mejorar el rendimiento, como el nivel de actividad física, la edad, o factores genéticos.

Predicciones

Usando la ecuación ajustada del modelo ($\boldsymbol{Y = 57.815 + 0.00428X}$), podemos realizar predicciones para valores específicos de consumo calórico (**X**):

1. Para X = 3000:

$$\boldsymbol{Y = 57.815 + 0.00428(3000) = 70.655\, kg}$$

2. Para X = 2500:

$$\boldsymbol{Y = 57.815 + 0.00428(2500) = 68.515\, kg}$$

3. Para X = 3500:

$$\boldsymbol{Y = 57.815 + 0.00428(3500) = 72.795\, kg}$$

Estas predicciones reflejan cómo el modelo estima el peso corporal en función del consumo calórico.

Conclusiones

- **Influencia de las variables independientes**: El consumo calórico (X) tiene un efecto positivo en el peso corporal (Y). Sin embargo, el coeficiente de regresión indica que este efecto es moderado.
- **Utilidad del modelo**: El modelo es útil para obtener una estimación general de la relación entre consumo calórico y peso corporal, pero su capacidad predictiva no es perfecta ($\boldsymbol{R^2 = 0.5034}$). Esto significa que hay otros factores importantes que el modelo no considera.
- **Errores y limitaciones**: El error promedio (RMSE y MAE) es significativo, lo que muestra que las predicciones pueden no ser precisas en todos los casos.

Recomendaciones:

- **Agregar más variables al modelo**: Incluir variables como edad, nivel de actividad física, metabolismo, género, y composición corporal podría mejorar significativamente el rendimiento del modelo.
- **Ampliar el tamaño de la muestra**: Un tamaño de muestra mayor podría proporcionar un modelo más robusto y confiable.
- **Análisis de residuos**: Realizar un análisis más profundo de los residuos para asegurarse de que cumplan los supuestos del modelo de regresión lineal (distribución normal y sin patrones sistemáticos).