

```
import pandas as pd
import statsmodels.api as sm
```

```
ruta = "Resultados encuesta.xlsx"
df = pd.read_excel(ruta)
```

```
df.head()
```

	¿Cuántas horas sueles dedicar al estudio a la semana? (en media)	¿En una escala del 1 al 10, qué tan satisfecho/a estás con tu rendimiento académico del curso anterior?	¿Cuántas horas dedicas a actividades extraescolares a la semana? (ej. deporte, baile, curso de fotografía, etc.)	De las horas anteriores, ¿cuántas serían dedicadas al deporte?	¿Cuántas horas dedicas a ver series a la semana?	¿Cuántas horas duermes en promedio cada noche?	¿Qué tipo de dispositivo utilizas más para estudiar?	¿Cuál es tu forma de transporte principal para llegar a la universidad? (Ej. Transporte público, coche, bicicleta, caminando)	¿Hasta qué punto te consideras de humanidades o ciencias? (3 sería igual de una cosa que de otra)	Nota media selectividad
0	9.0	8	4.0	2.0	3.0	7.5	portátil	metro y caminando	4	Na
1	3.0	9	5.0	0.0	1.0	6.0	portátil	Renfe	5	Na
2	4.0	10	6.5	4.0	3.0	6.0	portátil	coche	4	Na
3	5.0	9	3.0	0.0	3.0	7.0	portátil	bus	2	Na
4	12.0	8	0.0	0.0	5.0	7.0	portátil	metro y caminando	4	Na

```
# Para limpiar las columnas y renombrar las que voy a usar (me daba algún error en los próximos códigos al usar la frase completa)
df.columns = df.columns.str.strip()
df = df.rename(columns={
    "De las horas anteriores, ¿cuántas serían dedicadas al deporte?": "horas_deporte",
    "¿Cuántas horas duermes en promedio cada noche?": "horas_sueno",
    "¿Cuántas horas duermes en promedio cada noche? ": "horas_sueno" # por si tiene espacio final
})
```

He escogido las variables "De las horas anteriores, ¿cuántas serían dedicadas al deporte?" y "¿Cuántas horas duermes en promedio cada noche?". En mi caso la variable explicativa es el deporte y la explicada las horas de sueño.

Basándome en las conclusiones de alguno de los estudios que presento en el siguiente párrafo, quiero comprobar si las personas que hacen más deporte podrían dormir MENOS horas al tener un sueño de mejor calidad.

El objetivo es ver si hay una relación lineal entre el número de horas de deporte semanal y el número de horas de sueño diario.

He encontrado varias tesis doctorales y estudios sobre si el hacer deporte de forma moderada (1 o más horas al día) mejora la calidad del sueño, evita somnolencia durante el día, etc. Presento algunas:

- [Thompson, C. \(2020\). *Impact of Sleep Duration, Sleep Quality, and Physical Activity on Adolescent Health Behaviors*. Mississippi State University.](#)
- [Wang, E., Boros, S., & Shapiro, C. M. \(2021\). *The effect of physical exercise on sleep quality: A systematic review*. *Sleep and Biological Rhythms*, 19\(3\), 243–251.](#)
- [Ye, Y., Chen, J., & Zhang, L. \(2022\). *Effect of physical exercise on sleep quality of college students: Chain intermediary effect of mindfulness and ruminative thinking*. *Frontiers in Psychology*, 13, 987537.](#)

La pregunta del modelo es:

¿El número de horas de deporte que una persona hace a la semana explica el número de horas que duerme al día?

Expresión matemática de mi modelo:

Horas de sueño al día = $\beta_0 + \beta_1 \cdot (\text{Horas de deporte a la semana}) + \varepsilon$

Donde:

Horas de sueño al día → variable dependiente (Y)

Horas de deporte a la semana → variable independiente (X)

$\beta_0 \rightarrow$ intercepto o valor medio de horas de sueño cuando no se hace deporte

$\beta_1 \rightarrow$ cambio medio en las horas de sueño por cada hora adicional de deporte a la semana

$\varepsilon \rightarrow$ término de error (otros factores que influyen en el sueño)

Aplicación del método de los mínimos cuadrados y la recta estimada:

```
# variables
X = df["horas_deporte"]
Y = df["horas_sueno"]

#  $\beta_0$ 
X = sm.add_constant(X)

# Ajuste del modelo de regresión lineal simple
modelo = sm.OLS(Y, X).fit()
print(modelo.summary())

# Coeficientes estimados
beta0 = modelo.params["const"]
beta1 = modelo.params["horas_deporte"]

print(f"La recta estimada es: Horas de sueño al día = {beta0:.3f} + {beta1:.3f} * (Horas de deporte a la semana)")
```

```

                        OLS Regression Results
=====
Dep. Variable:          horas_sueno      R-squared:                0.017
Model:                  OLS              Adj. R-squared:          0.009
Method:                 Least Squares    F-statistic:              2.237
Date:                  Fri, 10 Oct 2025  Prob (F-statistic):      0.137
Time:                  11:28:12          Log-Likelihood:          -209.44
No. Observations:      133              AIC:                    422.9
Df Residuals:          131              BIC:                    428.7
Df Model:               1
Covariance Type:       nonrobust
=====
                    coef    std err          t      P>|t|      [0.025    0.975]
-----
const                7.0966     0.157    45.284     0.000     6.787     7.407
horas_deporte         0.0314     0.021     1.496     0.137    -0.010     0.073
=====
Omnibus:                 91.908    Durbin-Watson:           2.022
Prob(Omnibus):           0.000    Jarque-Bera (JB):        772.657
Skew:                   -2.308    Prob(JB):                1.66e-168
Kurtosis:                13.868    Cond. No.                11.6
=====

```

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
La recta estimada es: Horas de sueño al día = 7.097 + 0.031 * (Horas de deporte a la semana)

La recta estimada es: Horas de sueño al día = 7.097 + 0.031 * (Horas de deporte a la semana)

Dibujo de la recta estimada sobre su diagrama de dispersión:

```
import matplotlib.pyplot as plt
import numpy as np

# Variables
X_var = df["horas_deporte"]
Y_var = df["horas_sueno"]

# Coeficientes del modelo
beta0 = 7.097
beta1 = 0.031

# Puntos de la recta estimada
x_vals = np.linspace(X_var.min(), X_var.max(), 100)
y_vals = beta0 + beta1 * x_vals

# Diagrama de dispersión
plt.scatter(X_var, Y_var, color="blue", label="Datos observados")

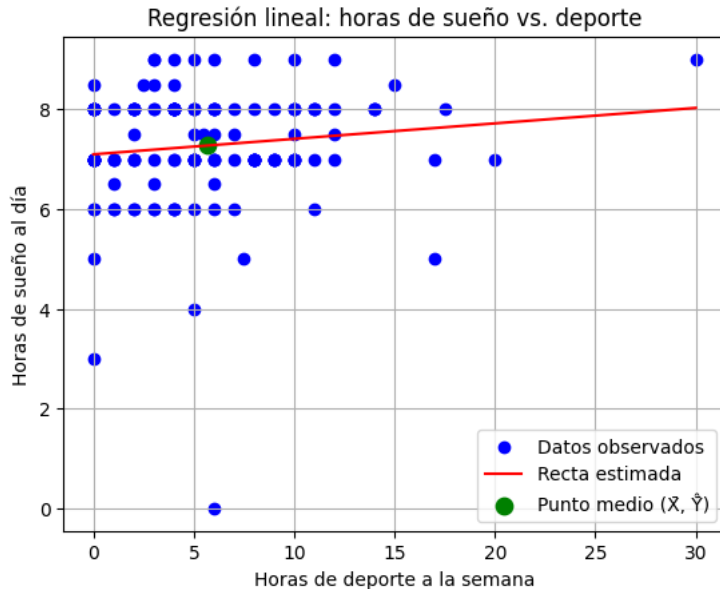
# Recta estimada
plt.plot(x_vals, y_vals, color="red", label="Recta estimada")

# Punto de medias (media de X, media de Y)
```

```
plt.scatter(X_var.mean(), Y_var.mean(), color="green", s=80, label="Punto medio ( $\bar{X}$ ,  $\bar{Y}$ )")

plt.xlabel("Horas de deporte a la semana")
plt.ylabel("Horas de sueño al día")
plt.title("Regresión lineal: horas de sueño vs. deporte")
plt.legend()
plt.grid(True)
plt.show()

# Punto medio
print(f"La recta pasa con toda seguridad por el punto ( $\bar{X}$ ,  $\bar{Y}$ ) = ({X_var.mean():.3f}, {Y_var.mean():.3f})")
```



La recta pasa con toda seguridad por el punto ($\bar{X}$, $\bar{Y}$) = (5.654, 7.274)

La gráfica muestra una relación ligeramente positiva entre las horas de deporte a la semana y las horas de sueño al día. La recta roja indica que, a medida que aumenta el tiempo dedicado al deporte, el sueño tiende a incrementarse ligeramente, aunque el efecto es pequeño y los puntos están muy dispersos, lo que sugiere una relación débil. El punto verde representa las medias de ambas variables, por donde la recta pasa necesariamente.

En conjunto, se observa que hacer más deporte podría asociarse con dormir un poco más, pero la influencia del deporte sobre el sueño diario es mínima y no se están teniendo en cuenta otros factores como el estrés, horarios, estudios, etc.

Yo pretendía encontrar una relación entre pocas horas de sueño y mucho deporte a la semana, lo cual en un principio puede sonar ilógico, pero varios estudios como los mencionados anteriormente han demostrado que el deporte mejora la calidad del sueño, no aumenta necesariamente la cantidad. Al tratarse de un estudio con exclusivamente dos variables y sin tener en cuenta otros factores, no nos sorprende el resultado.

Vamos a predecir las horas de sueño al día para una persona que hace 10 horas de deporte a la semana

```
x0 = 10

x_nuevo = pd.DataFrame({'const': [1], 'horas_deporte': [x0]})
print(x_nuevo)

# Intervalo de confianza al 99% de los coeficientes del modelo
int_conf_99 = modelo.conf_int(alpha=0.01)
print(int_conf_99)

# Predicción puntual y su intervalo de confianza
predicciones = modelo.get_prediction(x_nuevo)
predicciones = predicciones.summary_frame(alpha=0.01)

y_pred = predicciones["mean"][0]
ic_inf = predicciones["obs_ci_lower"][0]
ic_sup = predicciones["obs_ci_upper"][0]

print(f"\nPredicción puntual para {x0} horas de deporte/semana: {y_pred:.3f} horas de sueño/día")
print(f"IC del 99% para la observación individual: [{ic_inf:.3f}, {ic_sup:.3f}]")
```

```
const  horas_deporte
0      1             10
      0             0    1
const    6.686997  7.506252
```

```
horas_deporte -0.023513  0.086409
```

```
Predicción puntual para 10 horas de deporte/semana: 7.411 horas de sueño/día  
IC del 99% para la observación individual: [4.312, 10.510]
```

La predicción puntual del modelo solo es válida para valores de “horas de deporte a la semana” que estén dentro del rango observado en la muestra. Si los datos de la encuesta muestran valores entre 0 y 30 horas de deporte semanal, las predicciones solo pueden considerarse fiables dentro de ese intervalo; fuera de él, el modelo estaría extrapolando y los resultados no serían confiables.