Clase 5
Regresión Lineal Múltiple I

Análisis Avanzado de Datos

Gabriel Sotomayor

Revisión de la prueba

1. ¿Cuál de las siguientes afirmaciones sobre la correlación es FALSA?

a. La correlación es simétrica, lo que significa que no distingue entre variables explicativas y de respuesta.

b. Valores cercanos a +1 o -1 indican una mayor fuerza de la relación.

c. Un valor de correlación cercano a 0 indica una relación lineal débil o nula entre las variables.

d. La correlación siempre describe con precisión la relación entre variables, incluso si la relación no es lineal.

e. El valor de la correlación no cambia si se modifican las unidades de medida de las variables.

Revisión de la prueba

1. ¿Cuál de las siguientes afirmaciones sobre la correlación es FALSA?

a. La correlación es simétrica, lo que significa que no distingue entre variables explicativas y de respuesta.

b. Valores cercanos a +1 o -1 indican una mayor fuerza de la relación.

c. Un valor de correlación cercano a 0 indica una relación lineal débil o nula entre las variables.

d. **La correlación siempre describe con precisión la relación entre variables, incluso si la relación no es lineal.**

e. El valor de la correlación no cambia si se modifican las unidades de medida de las variables.

Revisión de la prueba

2. El R^2 o coeficiente de determinación da cuenta de:

a. La proporción de la variabilidad total en la variable dependiente que el modelo NO puede explicar.

b. La relación causal entre una variable independiente y una variable dependiente.

c. La precisión con la que el modelo puede predecir valores futuros de la variable independiente.

d. La diferencia entre los valores observados y los valores predichos en un modelo de regresión.

e. El porcentaje de la varianza en la variable dependiente que puede ser explicado por la varianza en la variable independiente.

Revisión de la prueba

2. El R^2 o coeficiente de determinación da cuenta de:

a. La proporción de la variabilidad total en la variable dependiente que el modelo NO puede explicar.

b. La relación causal entre una variable independiente y una variable dependiente.

c. La precisión con la que el modelo puede predecir valores futuros de la variable independiente.

d. La diferencia entre los valores observados y los valores predichos en un modelo de regresión.

e. **El porcentaje de la varianza en la variable dependiente que puede ser explicado por la varianza en la variable independiente.**

Revisión de la prueba

3. ¿Cuál es la interpretación correcta de un coeficiente beta (b) o pendiente en un modelo de regresión lineal simple?

a. La cantidad de varianza en la variable independiente (X) que puede ser explicada por la variable dependiente (Y).

b. El valor promedio de la variable dependiente (Y) cuando la variable independiente (X) es cero.

c. El grado de asociación entre la variable dependiente (Y) y la variable independiente (X).

d. La cantidad de cambio en la variable dependiente (Y) cuando la variable independiente (X) aumenta en una unidad.

e. La fuerza de la relación entre dos variables, similar al coeficiente de correlación.

Revisión de la prueba

3. ¿Cuál es la interpretación correcta de un coeficiente beta (b) o pendiente en un modelo de regresión lineal simple?

a. La cantidad de varianza en la variable independiente (X) que puede ser explicada por la variable dependiente (Y).

b. El valor promedio de la variable dependiente (Y) cuando la variable independiente (X) es cero.

c. El grado de asociación entre la variable dependiente (Y) y la variable independiente (X).

d. **La cantidad de cambio en la variable dependiente (Y) cuando la variable independiente (X) aumenta en una unidad.**

e. La fuerza de la relación entre dos variables, similar al coeficiente de correlación.

Revisión de la prueba

4. ¿Cuál es la interpretación correcta del intercepto (a) en un modelo de regresión lineal simple?

a. El cambio esperado en la variable dependiente (Y) por cada unidad de aumento en la variable independiente

b. El valor esperado de la variable independiente (X) cuando la variable dependiente (Y) es cero.

c. El valor esperado de la variable dependiente (Y) cuando la variable independiente (X) es cero.

d. La cantidad de variación en la variable dependiente (Y) que puede ser explicada por el modelo.

e. La pendiente de la línea de regresión.

Revisión de la prueba

4. ¿Cuál es la interpretación correcta del intercepto (a) en un modelo de regresión lineal simple?

a. El cambio esperado en la variable dependiente (Y) por cada unidad de aumento en la variable independiente

b. El valor esperado de la variable independiente (X) cuando la variable dependiente (Y) es cero.

c. **El valor esperado de la variable dependiente (Y) cuando la variable independiente (X) es cero.**

d. La cantidad de variación en la variable dependiente (Y) que puede ser explicada por el modelo.

e. La pendiente de la línea de regresión.

Revisión de la prueba

5. ¿Qué es un residuo en un modelo de regresión lineal?

a. Es la diferencia entre el valor predicho y el valor promedio de la variable dependiente (Y).

b. La suma de los errores cometidos por el modelo al predecir los valores de la variable independiente (X).

c. La diferencia entre el valor observado de la variable dependiente (Y) y el valor predicho ($\hat{y}$) por la recta de regresión.

d. El valor promedio de la variable dependiente (Y) en el modelo.

e. El valor predicho de la variable independiente (X) en el modelo.

Revisión de la prueba

5. ¿Qué es un residuo en un modelo de regresión lineal?

a. Es la diferencia entre el valor predicho y el valor promedio de la variable dependiente (Y).

b. La suma de los errores cometidos por el modelo al predecir los valores de la variable independiente (X).

c. **La diferencia entre el valor observado de la variable dependiente (Y) y el valor predicho ($\hat{y}$) por la recta de regresión.**

d. El valor promedio de la variable dependiente (Y) en el modelo.

e. El valor predicho de la variable independiente (X) en el modelo.

Revisión de la prueba

6. En el texto Esser, este plantea (al menos) cinco críticas a el enfoque que llama “Sociología de las Variables” el cual define como un enfoque en que se identifica una variable dependiente (explanandum) y se propone un conjunto de variables independientes (explanans) que podrían influir en ella, y la explicación se considera lograda cuando se puede atribuir la varianza de la variable dependiente a los efectos de las variables independientes. **Mencione y explique dos de las críticas planteadas por Esser a este enfoque.**

Problemas de la Sociología de las Variables

Incompletitud: Las relaciones entre variables establecidas en un contexto pueden no ser aplicables en otros, revelando la falta de leyes sociológicas generales y estables. La SV, al intentar explicar fenómenos sociales, frecuentemente se queda en explicaciones ad hoc, lo que limita su alcance y efectividad.

Significado Variable: Las variables estructurales pueden tener significados diferentes según el contexto cultural o social. Este problema de equivalencia funcional implica que las mismas variables no siempre tienen el mismo impacto en diferentes escenarios, lo que dificulta la creación de explicaciones universales.

Reduccionismo: Al reducir fenómenos sociales a simples relaciones entre variables, la SV pierde de vista la complejidad de las decisiones individuales y colectivas, y cómo estas influyen en los resultados sociales.

Problemas de la Sociología de las Variables

Interdependencia: Las estructuras sociales son procesos dinámicos donde las interacciones entre individuos y procesos son complejas. La SV no aborda adecuadamente cómo estas interdependencias afectan los resultados sociales, limitando la capacidad de la SV para explicar fenómenos complejos.

Falta de Sentido: La SV ignora el sentido subjetivo de las acciones individuales, centrándose solo en relaciones entre variables. Esto deja de lado la dimensión interpretativa crucial para una explicación sociológica completa, que considera las decisiones conscientes de los individuos.

Revisión de la prueba

a) Interpretación de los coeficientes del modelo (intercepto y beta de regresión):

Intercepto (-272.321): El valor del intercepto indica que, cuando una persona tiene cero años de escolaridad, su ingreso esperado sería -272.321 pesos. Aunque este valor no tiene sentido en términos prácticos (ya que no es posible tener ingresos negativos), sirve como punto de referencia en el modelo.

Beta para Años de escolaridad (77.114): Este coeficiente indica que, por cada año adicional de escolaridad, el ingreso esperado aumenta en 77.114 pesos. En otras palabras, la educación tiene un impacto positivo en los ingresos laborales.

b) Interpretación y evaluación del ajuste del modelo (R^2):

R^2 = 0.13: El valor de R^2 nos indica que el 13% de la variabilidad en los ingresos laborales puede ser explicada por los años de escolaridad. Aunque el modelo tiene una relación positiva entre la educación y los ingresos, el bajo valor de R^2 sugiere que hay muchos otros factores que también influyen en los ingresos y que no están considerados en este modelo.

Evaluaciones

Tarea 2: 9 de octubre

- Regresión lineal múltiple

Informe 1: 30 de Octubre

- Regresión lineal múltiple o regresión logística

Recordatorio de RLS

Recta de Regresión Mínimo-Cuadrática

La regresión lineal simple se utiliza para describir la relación entre dos variables, una independiente (explicativa) y una dependiente (respuesta), mediante una recta de regresión.

Fórmula General

La recta de regresión se expresa como:

$$\hat{y} = a + bx$$

Pendiente b: Indica el cambio promedio en la variable respuesta$y$por cada unidad de cambio en la variable explicativa x.
Ordenada en el origen a: Representa el valor predicho de$y$cuando x= 0. Sólo tiene significado estadístico cuando x toma valores cercanos a 0.

Coeficiente de Determinación R^2 y Varianza Residual

¿Qué es R^2? - R^2, conocido como el coeficiente de determinación, es una medida estadística que indica la proporción de la varianza en la variable dependiente Y que es explicada por la variable independiente X en un modelo de regresión.

- Se calcula como:

$$R^2 = 1 - \frac{\text{Varianza Residual}}{\text{Varianza Total de } Y}$$

  Donde:

  - **Varianza Residual:** Es la varianza de los residuos, es decir, la parte de Y que no es explicada por X.
  - **Varianza Total de Y**: Es la varianza de los valores observados de Y.

R^2 como Proporción Explicada

- R^2 indica cuánta de la varianza total de Y es explicada por el modelo.
- Un R^2 cercano a 1 sugiere que la mayor parte de la varianza de Y es explicada por X.
- Un R^2 cercano a 0 sugiere que el modelo no explica bien la varianza de Y, y la varianza residual es alta.

Interpretación Práctica de R^2: - Un R^2 de 0.18 indica que el 19% de la varianza en Y es explicada por X, mientras que el 81% restante es debido a factores no capturados por el modelo (varianza residual).

Objetivo de la sesión

Introducir el concepto de control estadístico y el uso de regresión lineal múltiple.

El problema del control estadístico.

El control estadístico consiste en ajustar los análisis para “controlar” el efecto de otras variables (covariadas) que podrían estar influyendo en la relación entre las variables de interés.

Ejemplo: En un estudio sobre diferencias salariales entre hombres y mujeres, las covariadas pueden incluir años de empleo o nivel educativo.

Distintas formas de control

Podemos controlar por otras variables que pueden influir en nuestros resultados a partir del diseño de nuestra investigación: A partir de asignación aleatoria en un experimento.

Por otro lado podemos controlar estadísticamente: Control Ajuste matemático que no requiere manipulación directa de datos o exclusión de casos. Es lo que comunmente tendremos que hacer en el contexto de estudios observacionales.

Ventajas del control estadístico

- No se manipulan participantes ni condiciones.
- No requiere excluir datos.
- Permite “mantener constantes” ciertas variables para observar el efecto “puro” de la variable independiente.
- Limitaciones: Requiere medición precisa de las covariadas y puede haber desacuerdo sobre qué variables controlar.

Regresión lineal múltiple

Introducción al Modelo de Regresión Múltiple

Un modelo de regresión múltiple examina la relación entre una **variable dependiente** y **varias variables independientes o predictores**.
La regresión múltiple permite **controlar otras variables** mientras se evalúa el efecto de una variable predictora específica.
Ejemplo: Si estudiamos la relación entre el ejercicio y la pérdida de peso, también podemos controlar la cantidad de alimentos consumidos para aislar su efecto.

Ecuación básica del modelo:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k + \epsilon$$

Donde Y es la variable dependiente, b_0 es la constante o intercepto, $X_1, X_2, \ldots X_k$ son las variables independientes, $b_1, b_2, \ldots b_k$ son los coeficientes de regresión, y ϵ es el error.

Asociación Parcial

La asociación parcial mide la relación entre dos variables manteniendo constantes otras variables.

Ejemplo: En un estudio sobre pérdida de peso, podemos medir la relación entre la ingesta de alimentos y la pérdida de peso, controlando la cantidad de ejercicio realizado.

Veamos cómo las asociaciones cambian al controlar variables adicionales.

Ejemplo

ID	Frecuencia de ejercicio (horas semanales promedio) X_1	Ingesta diaria promedio de alimentos (100s de calorías por encima del recomendado) X_2	Pérdida de peso semanal promedio (100s de gramos) Y
1	0	2	6
2	0	4	2
3	0	6	4
4	2	2	8
5	2	4	9
6	2	6	8
7	2	8	5
8	4	4	11
9	4	6	13

10	4	8	9
Promedio	**2**	**5**	**7.5**

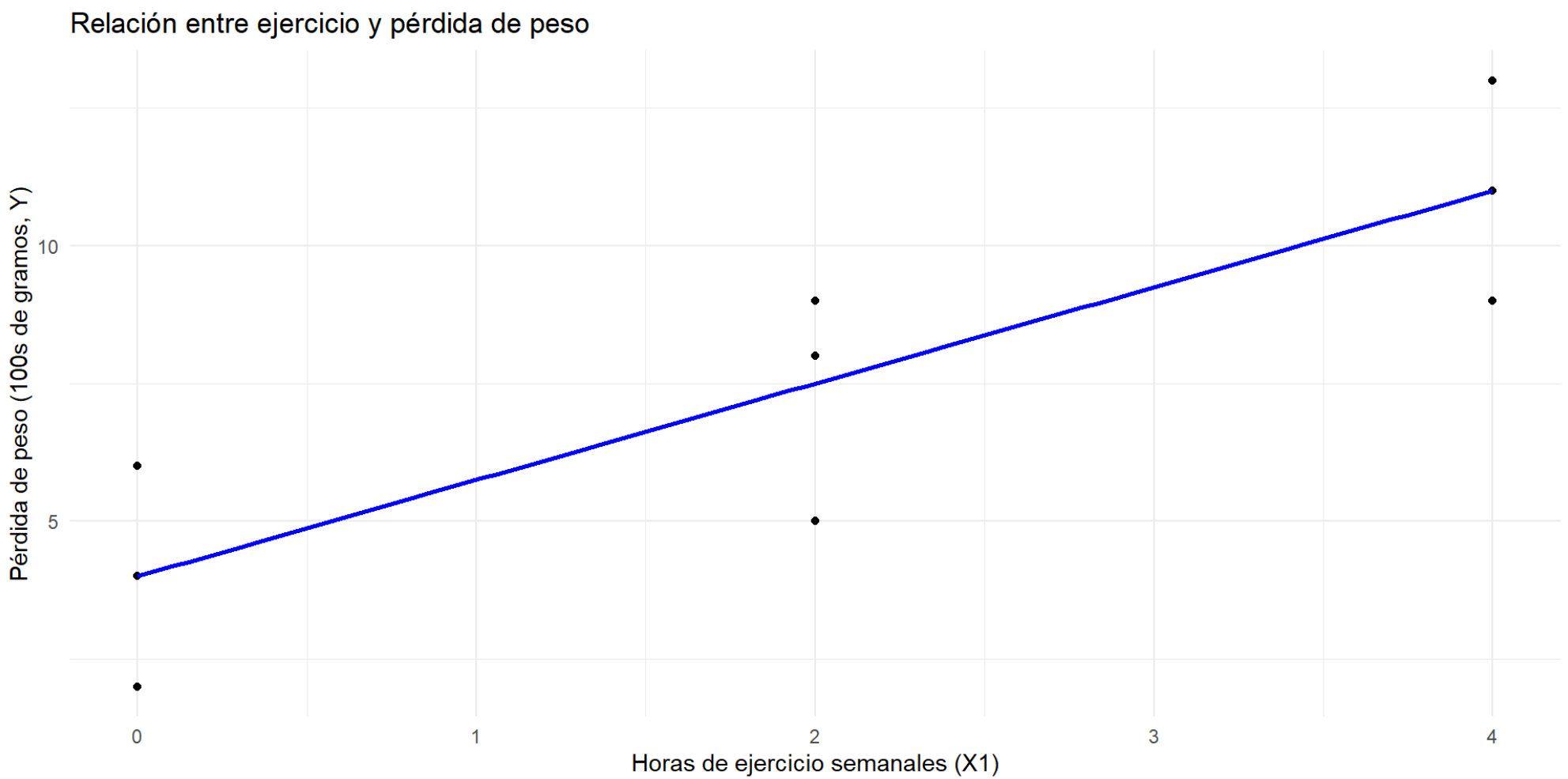
Correlación simple entre ejercicio y perdida de peso

La correlación entre ejercicio y pérdida de peso es positiva ($r_{X_1Y} = 0.864$).

Interpretación: Los participantes que hacen más ejercicio tienden a perder más peso.

Conclusión: Más ejercicio está asociado con mayor pérdida de peso.

Correlación simple entre ejercicio y perdida de peso



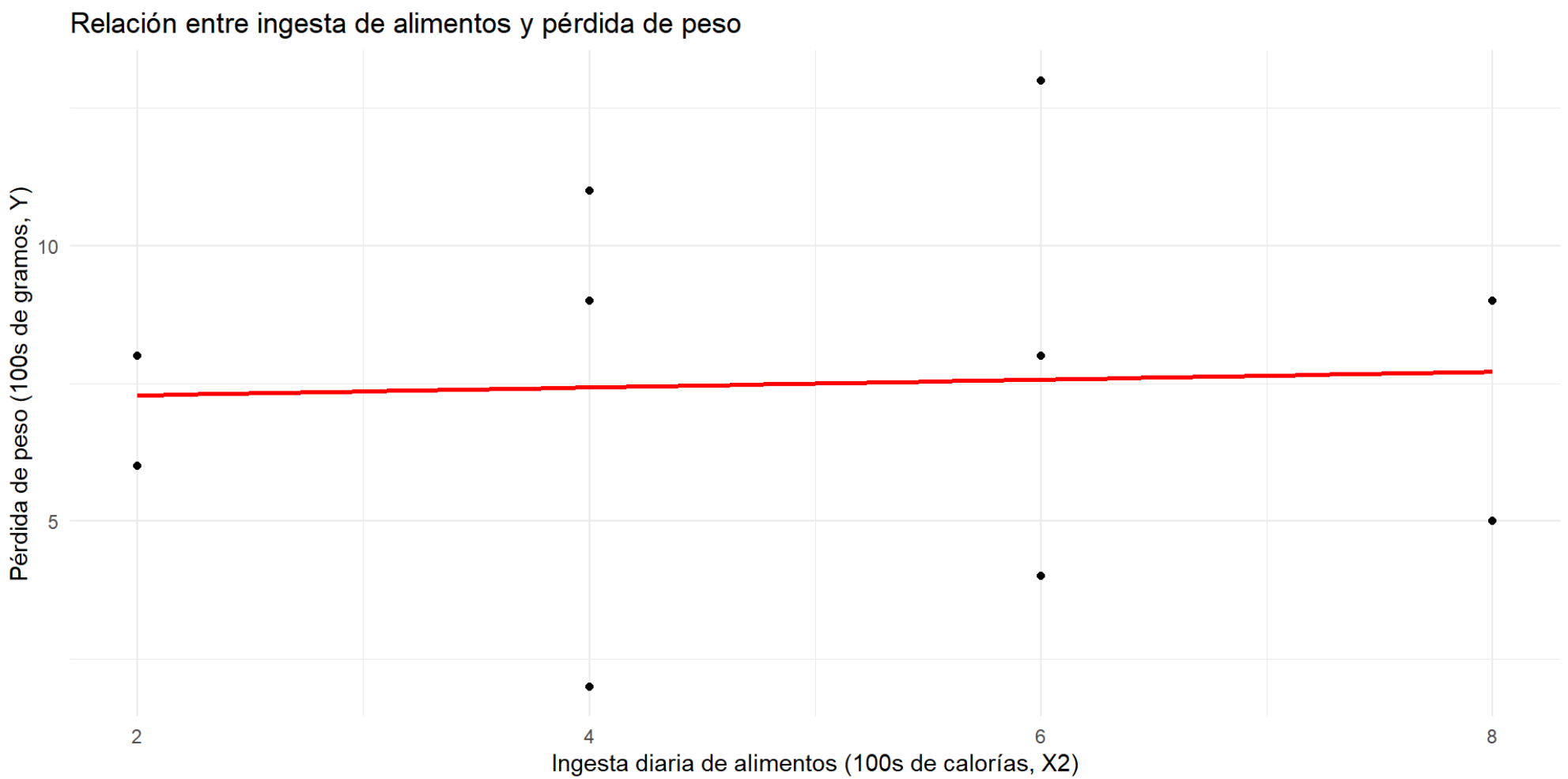
Correlación simple entre consumo de comida y Pérdida de Peso

La correlación entre ingesta de alimentos y pérdida de peso es pequeña y positiva ($r_{X_2Y} = 0.047$).

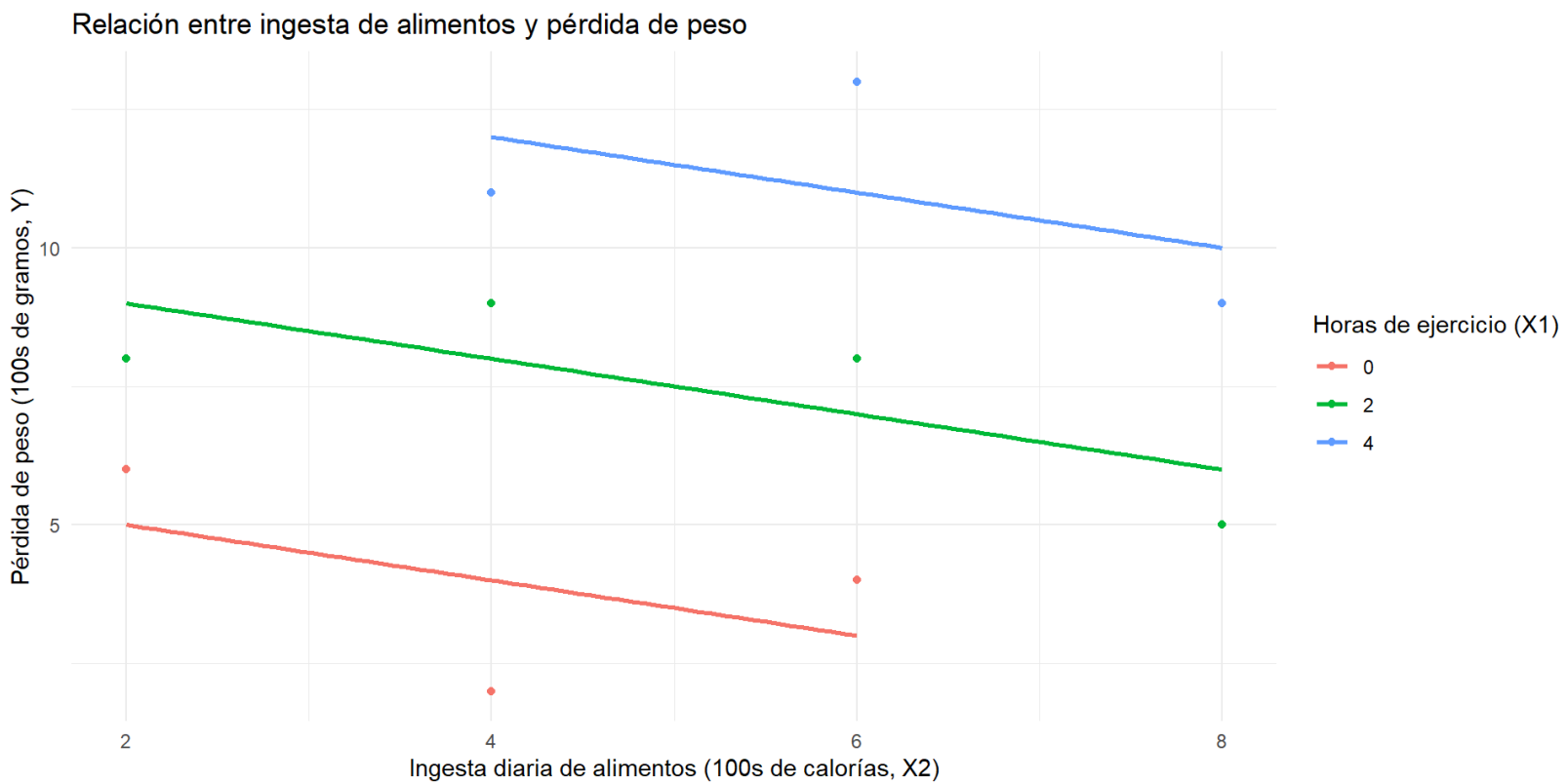
Contraintuitivo: Se esperaría que comer más implique perder menos peso, pero los datos sugieren lo contrario.

Pregunta: ¿Qué está pasando aquí? El ejercicio puede estar ocultando la verdadera relación entre comida y pérdida de peso.

Correlación simple entre consumo de comida y Pérdida de Peso



Relación entre consumo de comida y perdida de peso, controlando por ejercicio



Relación entre consumo de comida y perdida de peso, controlando por ejercicio

Al controlar el nivel de ejercicio, podemos ver la verdadera relación entre la ingesta de alimentos y la pérdida de peso.

Comer más realmente se asocia con perder menos peso, pero este efecto estaba oculto debido al impacto del ejercicio.

Modelo de Regresión Múltiple

Ecuación:

$$Y = 6 + 2X_1 - 0.5X_2$$

Interpretación:

- $b_0 = 6$: La pérdida de peso esperada (en cientos de gramos) para alguien que no hace ejercicio ni consume calorías extra.
- $b_1 = 2$: Por cada hora adicional de ejercicio semanal, se espera una pérdida de peso adicional de 200 gramos, manteniendo constante la ingesta de alimentos.
- $b_2 = -0.5$: Por cada 100 calorías extra consumidas, se espera perder 50 gramos menos, manteniendo constante el ejercicio.

El modelo permite aislar los efectos de cada predictor, controlando por los demás.

Modelo de Regresión Múltiple

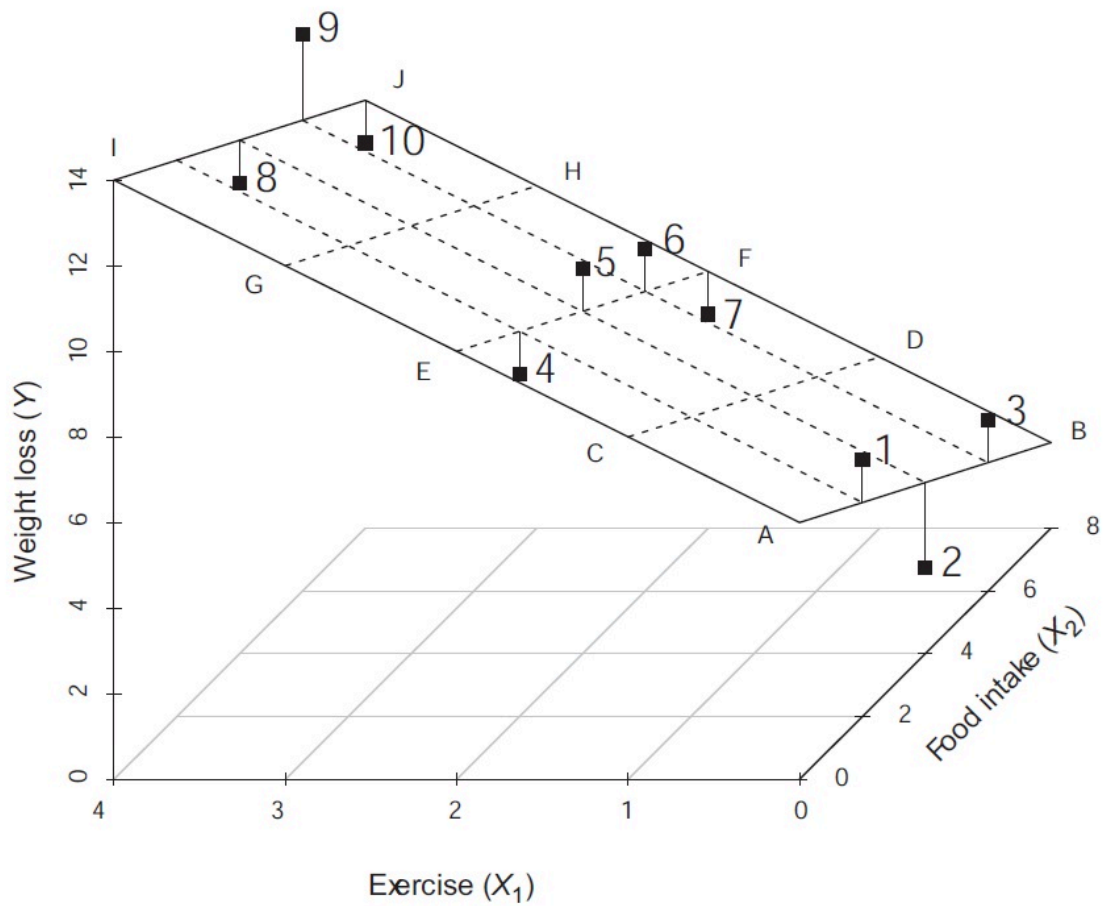


FIGURE 3.4. The data and the best-fitting plane.

Ajuste del modelo y residuos

Método de mínimos cuadrados: - El ajuste del modelo se realiza utilizando el método de mínimos cuadrados. - Este método busca minimizar la suma de los residuos al cuadrado, que son las diferencias entre los valores observados (Y) y los valores predichos ($\hat{Y}$) por el modelo.

Minimización de los residuos: - Los residuos (e_i) se calculan como la diferencia entre el valor observado y el valor predicho:

$$e_i = Y_i - \hat{Y}_i$$

- El objetivo del modelo es minimizar la suma de los residuos al cuadrado:

$$\sum (Y_i - \hat{Y}_i)^2$$

El uso de mínimos cuadrados garantiza que el plano de regresión ajustado se acerque lo más posible a los puntos de datos observados.