Investigación de Mercados II

Lección 7: La investigación longitudinal en marketing aplicando la regresión lineal

Contenidos:

1. Introducción a la regresión lineal
2. El modelo de regresión lineal
3. El modelo de regresión lineal con SPSS

1. Introducción a la regresión lineal

- El análisis de regresión **estudia la relación entre una variable a explicar** (dependiente) **con respecto a una o más variables explicativas** (independientes).
- Tiene como objetivo **determinar la ecuación matemática que relaciona a dichas variables para verificar hipótesis teóricas y/o poder predecir valores de la dependiente** realizando simulaciones con las independientes.
- A priori**, todas las variables** del modelo de regresión lineal **han de ser métricas,** aunque su posterior desarrollo ha permitido la **inclusión de categóricas.**

1. Introducción a la regresión lineal

- Matemáticamente, se puede expresar de la siguiente manera:

$$Y = f(X)$$

- La Y representa a la variable dependiente y la X al vector de variables explicativas (X_1, X_2, … X_k).
  - Si el número de variables explicativas es igual a uno, se habla de un **modelo lineal simple (MLS).**
  - Si el número de variables explicativas es mayor que uno, se habla de un **modelo lineal múltiple (MLM).**

1. Introducción a la regresión lineal

- La **aplicación de esta técnica** puede ser para:
  - **Series temporales**: observaciones sobre los valores que toma una variable en diferentes momentos de tiempo (mensual, anual, etc.).
  - **Datos de corte transversal**: observaciones de un conjunto de unidades (personas, familias, empresas, regiones, etc.).
  - **Información combinada**: datos agrupados que tienen elementos de series de tiempo y de corte transversal (por ejemplo los paneles).

1. Introducción a la regresión lineal

- El MLS se representa de la siguiente manera:

$$Y_i = \propto + \beta_1 x_1$$

  - Donde:
    - Y_i es la variable dependiente.
    - $\propto$ es la ordenada en el origen o término independiente.
    - β_1 es el coeficiente de x_1 o pendiente de la recta de regresión.
    - x_1 es la variable independiente.

1. Introducción a la regresión lineal

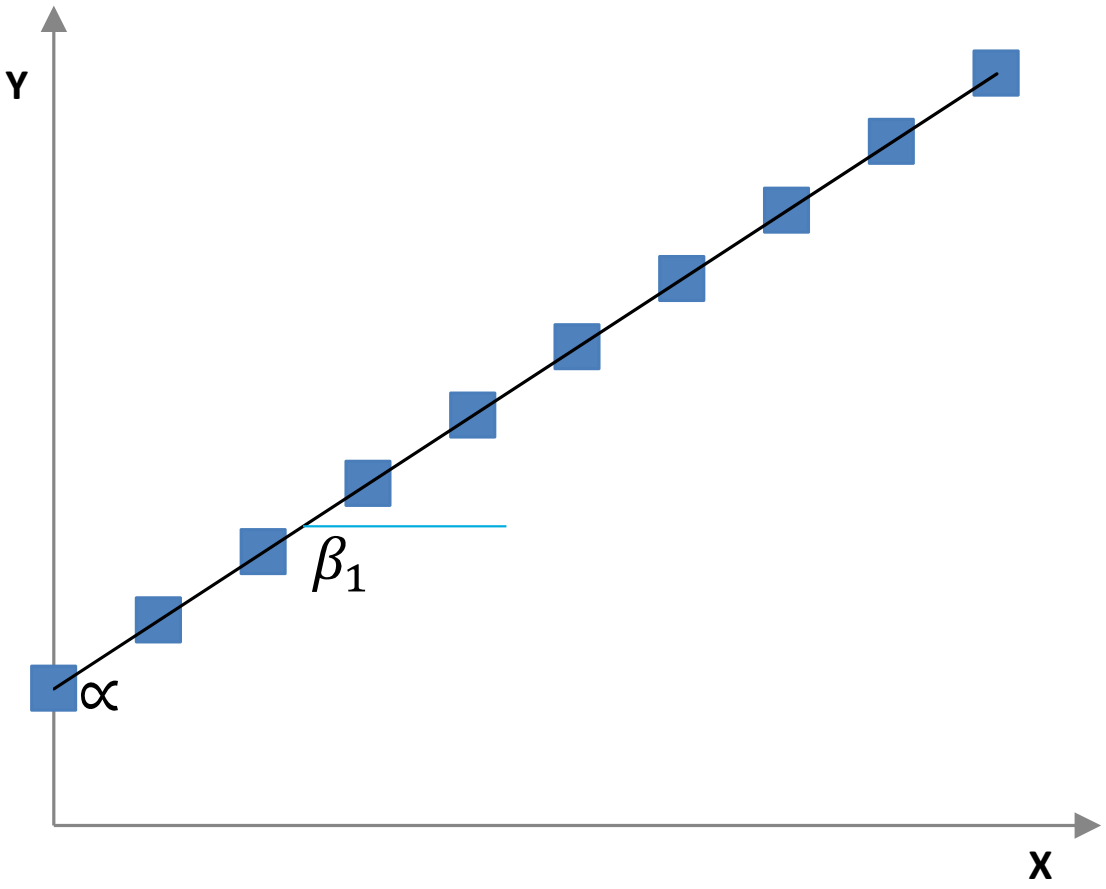
- Pero, la igualdad entre variables que representa el MLS rara vez ocurren. Más bien son aproximaciones donde se han cometido errores de especificación, lo que obliga a incluir en la ecuación un **término de perturbación aleatoria**:

$$Y_i = \propto + \beta_1 x_1 + u_i$$

1. Introducción a la regresión lineal

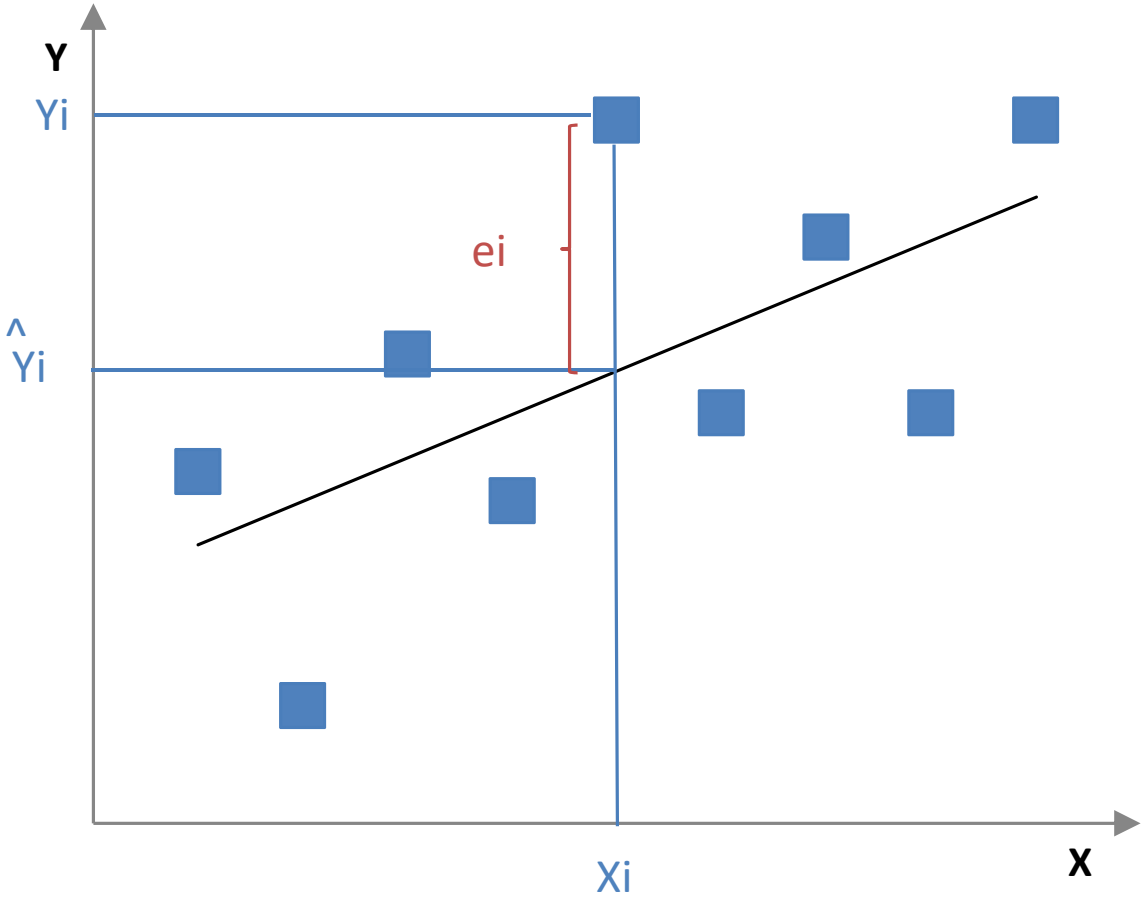
- La relación entre variables definida por el MLS se puede aproximar **gráficamente**:
  - Si las observaciones se sitúan **sobre una línea recta**, la relación entre X e Y será **exacta** y por tanto las ui serán cero.
  - En cambio, en el otro extremo, si las observaciones están totalmente dispersa formando una **nube de puntos**, la relación entre X e Y será completamente **estocástica o aleatoria**.

2. El modelo de regresión lineal



- Relación lineal exacta:
  - α será la ordenada en el origen.
  - β_1 será la pendiente de la recta.
  - Las perturbaciones son cero.

2. El modelo de regresión lineal



- En un **caso más realista**, nuestro objetivo será estimar el alfa y beta de la función que mejor aproxime a la nube de puntos de las observaciones.

- A las diferencias entre el valor observado y el estimado por el MLS se le denomina **error o residuo**: e_i

1. Introducción a la regresión lineal

- Existen diferentes **criterios o métodos** para ajustar la recta de regresión; siendo el más utilizado el de **mínimos cuadrados**:
  - Según este método, la mejor recta será aquella que **minimice la suma del cuadrado de los residuos**:

$$\min \sum e_i^2$$

- Generalizando el MLS, podemos llegar al **MLG (Modelo Lineal General):**

$$y_i = \propto + \beta_1 x_1 + \beta_2 x_2 + \quad + \beta_k x_k + u_i$$

2. El modelo de regresión lineal

Supuestos de Partida del MLG:

1. La relación funcional de las variables es de tipo lineal (*Hipótesis de Linealidad*). Es un supuesto teórico y no muy restrictivo, ya que este tipo de relación suele ser frecuente.
2. Las variables explicativas serán linealmente independientes entre ellas y con respecto a las perturbaciones del modelo (*Hipótesis de* ***Ausencia de Multicolinealidad Exacta***).
3. Las perturbaciones aleatorias son normales (*Hipótesis de Normalidad*). Esto implica:
   a) La Esperanza de las perturbaciones ha de ser cero: $E(u_i) = 0$
   b) **Homoscedasticidad** en las varianzas de las perturbaciones: $Var(u_i) = \sigma^2$
   c) **Ausencia de autocorrelación** entre las perturbaciones: $Cov(u_i; u_j) = 0$

2. El modelo de regresión lineal

1. **Supuesto de Multicolinealidad Exacta:**
   - El hecho de que exista multicolinealidad exacta entre las variables independientes **imposibilitará la obtención de los estimadores mínimo cuadrados**.
   - **Detección**:
     - **Significación del modelo**: si los coeficientes del modelo son no significativos individualmente y el modelo en su conjunto es significativo.
     - **Coeficientes de correlación**: si los coeficientes de correlación superan el 0,75.
       - Además, **si existen diferencias entre los coeficientes de correlación y los coeficientes de correlación parcial**.

2. El modelo de regresión lineal

1. **Supuesto de Multicolinealidad Exacta:**
   - **Detección**:
     - **Tolerancia:** valores próximos a 1 indicarán ausencia de multicolinealidad y próximos a 0, indicarán multicolinealidad.
     - **Factor de Inflación de la Varianza (FIV):** en ausencia de multicolinealidad valdrá 1, y valores superiores a 4 indicarán presencia de multicolinealidad; aunque a partir de 2 empiezan a considerarse problemáticos.
     - **Autovalores**: si son próximos a cero, indicarán un problema de multicolinealidad.
     - **Índice de Condición**: se suele considerar la presencia de multicolinealidad grave cuando el índice está por encima de 30; aunque por encima de 15 indica un posible problema de multicolinealidad.

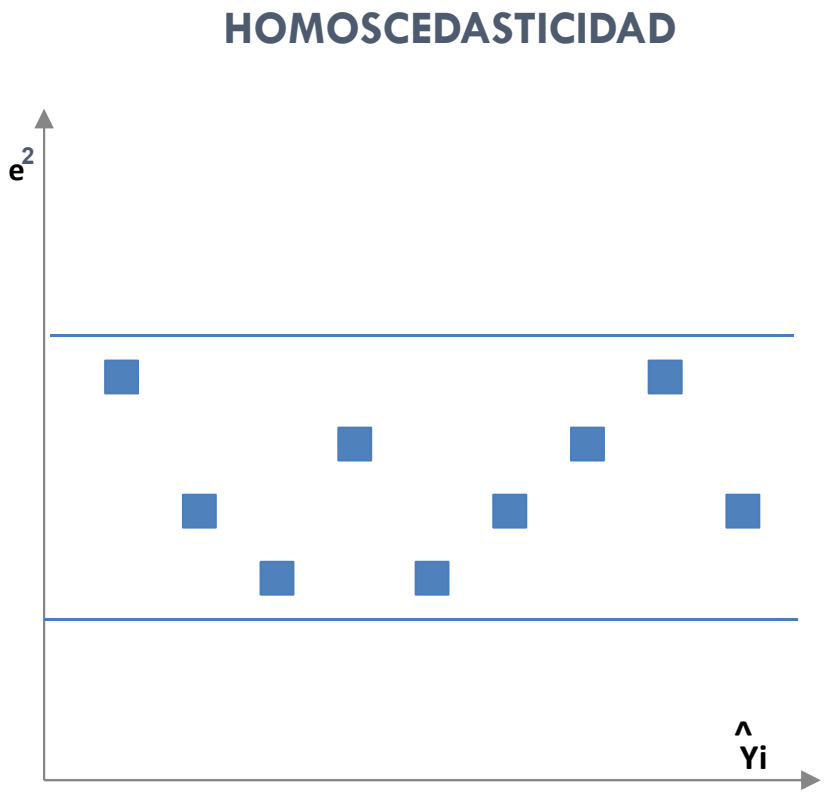
2. El modelo de regresión lineal

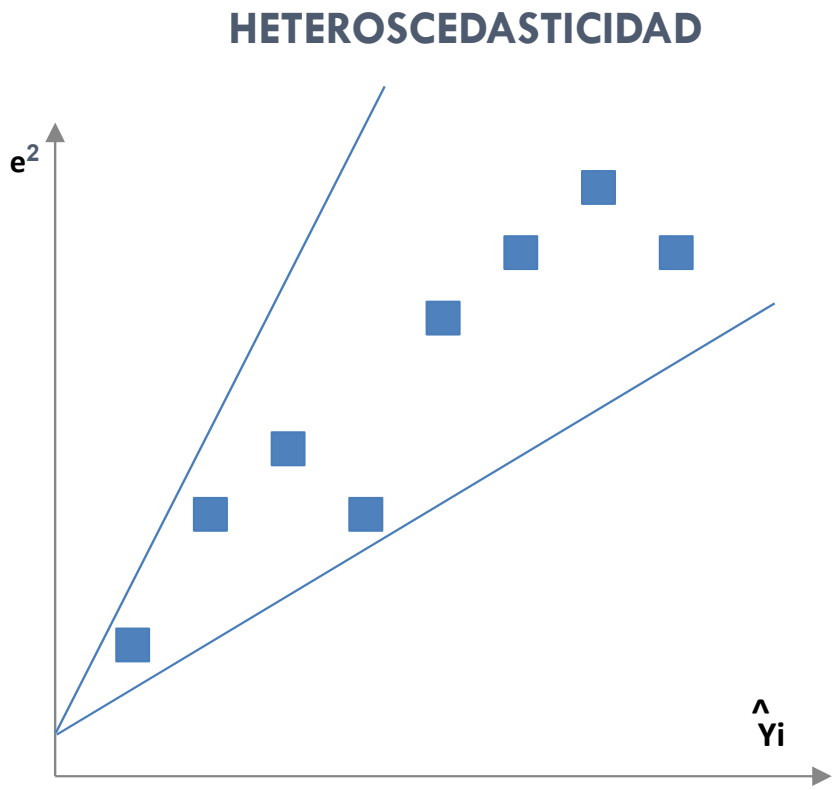
2. Supuesto de Homoscedasticidad de las Perturbaciones:

- **Detección**: el recurso más usado son las representaciones **gráficas de los residuos**.
  - En un eje se representan los residuos al cuadrado y en el otro eje los valores estimados de la variable dependiente.
  - Si se observa algún tipo de patrón sistemático, entonces existirá heteroscedasticidad.
  - Además, es conveniente realizar el **mismo gráfico para cada una de las variables explicativas**, identificando cuál de ellas genera la heteroscedasticidad.

2. El modelo de regresión lineal

2. Supuesto de Homoscedasticidad de las Perturbaciones:





2. El modelo de regresión lineal

3. **Supuesto de Ausencia de Autocorrelación entre Perturbaciones:**

   - Se suele deber a problemas de especificación del modelo y se dice que **existe** cuando la **covarianza para cada par de valores es distinta de cero.**

   - **Detección:** el test **más usado es el de Durbin-Watson ($0 \leq d \leq 4$):**

     - Si $d > 2$, autocorrelación positiva

     - Si $d < 2$, autocorrelación negativa.

     - Aunque **si $1,5 \leq d \leq 2,5$,** se suele asumir la ausencia de autocorrelación.

2. El modelo de regresión lineal

- Cuando se cumplen estas hipótesis, a los estimadores mínimo cuadrados, se les llama **mínimos cuadrados ordinarios (MCO), y en este caso presentan las siguientes características:**

  a) Son **insesgados**, es decir, la diferencia entre el verdadero valor del parámetro y el valor esperado del estimador es cero:

$$\beta - E(\hat{\beta}) = 0$$

  b) **Tienen varianza mínima** entre todos los estimadores lineales e insesgados, por lo que **se dice que es un estimador lineal, insesgado y óptimo (ELIO).**

  c) **Son consistentes**, ya que a medida que aumenta la muestra, **el estimador converge hacia el verdadero valor del parámetro**.

2. El modelo de regresión lineal

- El **modelo estimado por MCO**, se expresaría de la siguiente manera:

$$Y_i = \widehat{\beta_1} + \widehat{\beta_2} X_{2i} + \widehat{\beta_3} X_{3i} + \cdots + \widehat{\beta_k} X_{ki} + e_i$$

- **Donde los coeficientes serán la variación esperada que se produce en Y** (en las unidades de medida en las que venga dada dicha variable) **cuando se incrementa en una unidad la variable explicativa correspondiente**, **suponiendo que el resto de variables explicativas permanecen constantes.**

2. El modelo de regresión lineal

- El EMCO nos proporciona una **estimación puntual**, pero a veces interesa realizar adicionalmente una **estimación por intervalos**.
- La estimación por intervalos proporcionará **un intervalo dentro del cual se encontrará el verdadero valor del parámetro dado un nivel de confianza $(1-\alpha)$.**
- En la práctica el intervalo de estimación para un parámetro β_j viene dado por:

$$\hat{\beta}_j \pm \hat{\sigma}_{\hat{\beta}_j} t_{n-k,\alpha/2}$$

2. El modelo de regresión lineal

- Otro aspecto importante que nos proporciona el análisis de regresión es la **significatividad de los parámetros estimados**.
- Para verificar la **significatividad de un parámetro individualmente**, se usa el siguiente estadístico: $t = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}}$, que se distribuye como una t-Student con n-k grados de libertad, y contrasta la siguiente hipótesis:
  - H0: $\beta_j = 0$ [El parámetro no es significativo]
  - H1: $\beta_j \neq 0$ [El parámetro sí es significativo]

2. El modelo de regresión lineal

- Igualmente plantea si **el modelo es significativo en su conjunto**, esto es, si de manera conjunta el modelo explica o no las variaciones de la variable dependiente.
  - H0: $\beta_2 = \beta_3 = \cdots = \beta_k = 0$ [El modelo no es significativo]
  - H1: La Ho no se cumple [El modelo sí es significativo]
- Este tipo de prueba, en el modelo de regresión, se conoce como análisis de la varianza o tabla ANOVA. El estadístico de contraste es: $F = \frac{SCE/(k-1)}{SCR/(n-k)}$, que se distribuye con k-1 y n-k grados de libertad.

2. El modelo de regresión lineal

- Además, el análisis de regresión permite obtener una medida de la bondad del ajuste del modelo, denominada **coeficiente de determinación R^2**, que mide **la proporción de las variaciones de la variable dependiente que vienen explicadas por el modelo**.
  - Un inconveniente del R^2 es que tiende a seleccionar modelos con mayor número de variables independientes.
  - Por ello, para comparar modelos con la misma variable dependiente y distinto número de independientes se recurre al **coeficiente de determinación ajustado $\overline{R}^2$**.

2. El modelo de regresión lineal

- La última fase suele ser **realizar predicciones de la variable dependiente** y, se demuestra que el predictor lineal, insesgado y óptimo (PLIO) es el que se obtiene sustituyendo en la expresión del MLG los parámetros por el EMCO.

$$\hat{Y}_0 = \widehat{\beta_1} + \widehat{\beta_2} X_{20} + \widehat{\beta_3} X_{30} + \cdots + \widehat{\beta_k} X_{k0}$$

$$\hat{Y}_1 = \widehat{\beta_1} + \widehat{\beta_2} X_{21} + \widehat{\beta_3} X_{31} + \cdots + \widehat{\beta_k} X_{k1}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\hat{Y}_n = \widehat{\beta_1} + \widehat{\beta_2} X_{2n} + \widehat{\beta_3} X_{3n} + \cdots + \widehat{\beta_k} X_{kn}$$

2. El modelo de regresión lineal

- **Inclusión de variables categóricas en el modelo**. Para incluir este tipo de variables se utilizan **variables ficticias**:
  - Si la variable cualitativa es **dicotómica o binaria**, **no existen problemas** con su inclusión en el modelo.
  - Si la variable cualitativa tiene **más de dos categorías**, a la hora de incluirla en el modelo **solo se podrán introducir m-1 variables ficticias**, donde **m es el número de categorías que presenta dicha variable**.
- **Si no se tiene en cuenta esta regla, el modelo presentará multicolinealidad perfecta y no será posible su estimación por MCO**.

2. El modelo de regresión lineal

- **Inclusión de variables categóricas en el modelo**. Ejemplos:
  - Ejemplo **Variable Ficticia Dicotómica**: Género (m = 2)

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 Z_i \ldots + \beta_k X_{ki} + u_i$$

Donde Z_i puede tomar 0 si es hombre o 1 si es mujer.

  - Ejemplo **Variable Ficticia Dicotómica**: Estado Civil (m = 3)

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 Z_{1i} + \beta_5 Z_{2i} \ldots + \beta_k X_{ki} + u_i$$

Donde Z_{1i} es la variable soltero y toma 0 si no lo es y 1 si lo es.

Donde Z_{2i} es la variable casado y toma 0 si no lo es y 1 si lo es.