Investigación de Mercados II

Tema 8: El estudio del comportamiento del consumidor con la regresión logística binaria

Contenido

1. Introducción a la regresión logística binaria
2. El modelo de regresión logística binaria
3. La regresión logística binaria con SPSS

1.Introducción a la regresión logística binaria

- Ejemplo: *El director de marketing de una empresa de comunicación le interesa conocer como ciertas características socioeconómicas (estado civil, ingresos, nivel de estudios, edad y número de hijos) influyen en un individuo para contratar un nuevo servicio de TV por cable o no.*

- Esta cuestión se puede plantear construyendo un modelo que describa como algunas **variables independientes (continuas o categóricas)** afectan a una **variable dependiente binaria, dicotómica o dummy.**

1.Introducción a la regresión logística binaria

- En primera instancia, podríamos pensar en un **análisis discriminante**, pero el hecho de que **existan variables independientes categóricas viola el supuesto de normalidad multivariante.**
- Otra opción podría ser plantearnos un modelo de **regresión lineal**, pero la **variable dependiente**, al ser discreta- dicotómica, **no se va a distribuir linealmente.**
- Para solventar las limitaciones de ambos modelos, sería necesario utilizar un modelo que fuese **no lineal pero monótono, creciente y acotado entre 0 y 1.**
- **El modelo de regresión logística** es una alternativa que **cumple dichos**

1.Introducción a la regresión logística binaria

- En estos casos, la variable dependiente se va a distribuir como:

$$y_i \sim \mathcal{B}\,(1; p_i)\ ;\ \text{(Distribución de Bernoulli)}$$

  - Donde n es igual al número de ensayos.
  - Donde p_i es la probabilidad de que suceda una opción u otra en función del número de ingresos del cliente:
    - $P\,(Y = 1\ /X = x_i) = p_i$
    - $P\,(Y = 0\ /X = x_i) = 1 - p_i$

- En otras palabras, la **variable dependiente explica la probabilidad de que suceda una opción u otra**.

1.Introducción a la regresión logística

- Considerando solo una variable independiente, la regresión logística determina la pro **la probabilidad de que Y=1, condicionado a un valor de X de la siguiente manera**:

$$p_i = \frac{e^{\beta_0+\beta_1 X_1}}{1 + e^{\beta_0+\beta_1 X_1}}$$

$$(1 - p_i) = \frac{1}{1 + e^{\beta_0+\beta_1 X_1}}$$

1.Introducción a la regresión logística

- Como no se cumplen los supuestos del modelo de regresión lineal, las estimaciones por MCO no van a ser eficientes. Por ellos, el modelo de regresión logística utiliza el método de **estimación por máxima verosimilitud (EMV):**

$$\hat{p}_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_1}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X_1}}$$

$$(1 - \hat{p}_i) = \frac{1}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X_1}}$$

1.Introducción a la regresión logística

Además, es necesario hablar de **dos conceptos asociados** a este análisis:

1) **Odds o ventaja de que un suceso ocurra:** se define como el cociente entre la probabilidad de que ocurra un suceso y su probabilidad complementaria. Por tanto, **indica la preferencia de elegir la opción 1** de la variable respuesta frente a la opción 0.

$$\widehat{\Omega}_i = \frac{p_i}{(1-p_i)} = e^{\widehat{\beta}_0+\widehat{\beta}_1 X_1} \; ; \qquad (0 \leq \widehat{\Omega}_i \leq \infty)$$

2) **Logit o transformación logística de la ventaja:** que expresa en términos lineales la preferencia de elegir la opción 1 de la variable respuesta frente a la opción 0.

$$\text{Ln}\, \widehat{\Omega}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_1; \qquad (-\infty \leq \text{Ln}\, \widehat{\Omega}_i \leq \infty)$$

1.Introducción a la regresión logística

Entendido el modelo, la generalización del mismo nos lleva a **estimar la probabilidad de que una respuesta binaria ocurra P (Y= 1), en función a los valores de un conjunto de variables explicativas**:

$$\hat{p}_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_{1i}+\hat{\beta}_2 X_{2i}+ \ldots+\hat{\beta}_n X_{ni}}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X_1}}$$

O sus expresiones equivalentes:

$$\hat{\Omega}_i = \frac{p_i}{(1 - p_i)} = e^{\hat{\beta}_0+\hat{\beta}_1 X_{1i}+\hat{\beta}_2 X_{2i}+ \ldots+\hat{\beta}_n X_{ni}}$$

$$\text{Ln}\, \hat{\Omega}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_n X_{ni}$$

2. El Modelo de Regresión Logística

1. **Codificación de las variables independientes categóricas**: no es correcto introducirlas como si fueran variables continuas; **necesitan un proceso de adaptación:**
   - Ante la presencia de una **variable cualitativa con k niveles**, será necesario diseñar o **definir k-1 variables dicotómicas** (ficticias o de diseño). Cada una de ellas, representará un nivel o una categoría de la variable original.

- *Ejemplo: el responsable de marketing considera además que la predisposición de sus clientes a adquirir el nuevo producto depende no sólo de los ingresos, sino de otras variables como:*
  - *Edad: "menos de 40 años", "de 40 a 60 años" y "más de 60 años".*
  - *Propiedad de vivienda: "sí" y "no".*

2. El Modelo de Regresión Logística

1. **Codificación de las variables independientes categóricas**:
   - *Para el caso de la variable "Propiedad de la vivienda" no es necesario definir ninguna variable nueva porque es dicotómica.*
   - *Pero, para la variable "Edad", habría que definir k-1= 2 variables ficticias; por ejemplo: "EDAD1 (menos de 40 años)" y "EDAD2 (de 40 a 60 años)".*
     - *Si un individuo tiene menos de 40 años, puntuaría 1 en EDAD 1 y 0 en EDAD2.*
     - *Si un individuo tiene "de 40 a 60 años" puntuaría 0 en EDAD 1 y 1 en EDAD2.*
     - *Y si un individuo tiene "más de 60 años" puntuaría 0 en ambas.*
   - Este **método de codificación de variables ficticias** recibe el nombre de **INDICADOR**.

2. El Modelo de Regresión Logística

2. **Significación de los coeficientes de regresión**:
   - Ajustado el modelo, hay que **comprobar si las variables independientes están relacionadas significativamente** (tanto a nivel individual como conjuntamente) con la variable respuesta o dependiente:
     - Para **la significación individual** se formulan las siguientes hipótesis:
       - H0: $\beta_i = 0$
       - H1: La Ho no se cumple
     - Para contrastar esta hipótesis se utiliza **estadístico de Wald**:

2. El Modelo de Regresión Logística

2. **Significación de los coeficientes de regresión** :

- Para la **significación global** se formulan las siguientes hipótesis:
  - H0: $\beta_2 = \beta_3 = \cdots = \beta_k = 0$
  - H1: La Ho no se cumple
- Para contrastar esta hipótesis se utiliza **estadístico G o Prueba de la Razón de Verosimilitud (Chi- Cuadrado para el Modelo en SPSS)**:

2. El Modelo de Regresión Logística

3. **Bondad del ajuste**:

   a) **Prueba de Hosmer- Lemeshow**, que formula has siguientes hipótesis:

      - H0: El modelo ajusta bien los datos observados
      - H1: H0 no es cierta

   b) **Histograma de probabilidades estimadas o Gráfico de Clasificación:** si el modelo distingue acertadamente, los casos para los que se ha observado que ocurre el fenómeno (Y= 1) deberían estar situados a la derecha del punto de corte (por defecto 0,5) y viceversa.

2. El Modelo de Regresión Logística

3. **Bondad del ajuste**:

   c) **Tabla de clasificación,** que en su diagonal principal recoge el total de casos bien clasificados. Además, nos devuelve **tres indicadores**:

   - **Tasa de aciertos**: casos bien clasificados/ total de casos
   - **Especificidad**: negativos correctos/ negativos observados.
   - **Sensibilidad**: positivos correctos/ positivos observados

   d) **Medidas similares al R^2** como el R2 de Cox y Snell y el R2 de Nagelkerke. Se tratan de variantes del primero y se interpretan de la misma manera.

2. El Modelo de Regresión Logística

4. **Interpretación de los resultados**:

- Variables **independientes dicotómicas**: será una medida de asociación que indica cuanto más probable (o improbable) es que se presente el suceso que se está investigando (Y = 1) entre aquellos sujetos para los que Xi valga 1 que entre aquellos otros con Xi igual a 0.
- *En el EJEMPLO 1: para la variable VIVIENDA, el valor estimado del exponente de Beta es de 35,796. Esto significa que, manteniendo constantes el resto de las variables, el mostrarse favorable a la adquisición del nuevo producto es 35,796 veces más probable que ocurra entre los sujetos que poseen vivienda propia (VIVIENDA= 1) que entre los que no la poseen (VIVIENDA= 0).*

2. El Modelo de Regresión Logística

4. **Interpretación de los resultados**:

- Variables **independientes categóricas con más de dos opciones de respuesta.** Su interpretación va a ser similar que para las independientes dicotómicas, salvo que la comparación se realiza respecto a la categoría de referencia (en este caso “Más de 60 años”):
- *En el EJEMPLO 1:*
  - *Para la variable EDAD1: el estar dispuesto a adquirir el nuevo producto financiero es 0,047 veces más probable que ocurra teniendo menos de 40 años que si el sujeto tiene más de 60 años (categoría de referencia).*
  - *Para la variable EDAD2: el estar dispuesto a adquirir el nuevo producto financiero es 6,245 veces más probable que ocurra teniendo entre 40 y 60 años que si el sujeto tiene más de 60 años.*

2. El Modelo de Regresión Logística

4. **Interpretación de los resultados**:

- Variables **independientes continuas.** Se interpreta de la siguiente manera: ante un cambio unitario en la escala de medida de la variable independiente, la ventaja de la opción 1 de la variable dependiente se incrementará (o disminuirá) en un factor igual al exponente de Beta:
- *En el EJEMPLO 1: la variable INGRESOS es de naturaleza continua y una vez estimado el modelo el Exponente de Beta es 1,207. Esto significa que un incremento de 1.000 euros en los ingresos (la unidad de medida de INGRESOS es en miles de euros) provocará un incremento multiplicativo de 1,207 veces en la ventaja de la opción 1 de la variable dependiente.*