Maestría en Marketing

Métodos y técnicas de análisis cuantitativo y cualitativo

Extensiones del modelo lineal

• Referencia: G. Caps 6 y 15. D Cap 1.

Otoño 2004

Modelo lineal con *k* variables

• Utilizando *k* variables explicativas, el modelo queda:

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

• Seguiremos suponiendo que los residuos tienen distribución $N(0, \sigma^2)$.

• Como antes, suponemos cada X no pueden tomar el mismo valor para todas los observaciones. También debemos suponer que las variables explicativas (ni una combinación lineal de ellas) no están perfectamente correlacionadas (No Multicolinealidad)

Bondad de Ajuste

• Como antes, podemos calcular

$$R^2 = \frac{SCE}{SCT}$$

• Problema: se puede mostrar que el R^2 crece (por construcción) a medida que aumentamos la cantidad de variables explicativas.

• Para salvar de este problema, definimos el R^2 ajustado

$$R_A^2 = 1 - \left(1 - R^2\right)\frac{N-1}{N-k}$$

Test de Hipótesis

• El test de significatividad individual (test t) sigue siendo

$$t = \frac{\hat{\beta} - 0}{\sqrt{Var(\hat{\beta})}}$$

• En este marco, también realizaremos un test de significatividad global (test F), donde las hipótesis son:

$$H_0 : \beta_2 = \beta_3 = ... = \beta_k 0$$
$$H_A : \beta_2 \neq 0 \ o \ \beta_3 \neq 0 \ o ... \ o \ \beta_k \neq 0$$

• Y el estadístico es:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(N-k)}$$

Omisión de variables relevantes

• Supongamos que el verdadero modelo es:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

• Pero nosotros estimamos omitiendo la variable X_3, es decir:

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i$$

• Se puede mostrar que, cuando hacemos esto, el estimador es sesgado

$$E(\hat{\beta}_2) \neq \beta_2$$

• Salvo que:

  • La variable X_3 no sea relevante para explicar a Y (β_3=0)

  • Las variables X_2 y X_3 no estén relacionadas $Corr(X_2, X_3)=0$

Multicolinealidad cercana

• Sabemos que no puede ser que dos variables sean tales que

$Corr(X_2, X_3)=1$ o $Corr(X_2, X_3)=-1$

• Pero que pasa si la correlación no es exacta pero es casi exacta?

• Se puede mostrar que, en este caso, la varianza de las estimadores será más alta, lo que producirá que sea menos probable rechazar la hipótesis de insignificatividad individual.

• Lo que esta pasando es que ambas variables capturan casi el mismo, efecto; por lo que no podemos identificar el efecto de ambas por separado.

Modelo lineal en parámetros

• Hasta ahora hemos asumidos que existe una relación lineal entre las variables. La pregunta es, hasta dónde podemos relajar este supuesto y seguir utilizando el método de mínimos cuadrados.

• Podremos seguir utilizando el método siempre que el modelo sea lineal en los parámetros. Esto es, podemos llegar hasta el punto donde

$$g_1(Y_i) = \beta_1 + \beta_2 g_2(X_{2i},\ldots,X_{ki}) + \ldots + \beta_k g_k(X_{2i},\ldots,X_{ki}) + u_i$$

• Lo que cambiará es la interpretación de cada coeficiente según sean las distintas funciones *g( )*. Hasta ahora, los coeficientes los interpretabamos como

$$\frac{dY}{dX} = \beta$$

Modelo lineal en parámetros

• Modelo **logarítmico**:

$$\log(Y_i) = \beta_1 + \beta_2 \log(X_{2i}) + u_i$$

• En este caso, el coeficiente se intrepreta como:

$$\beta = \frac{dY / Y}{dX / X}$$

• Esto es, cómo cambia porcentualmente *Y* ante un cambio de 1% en la variable *X*. Este tipo de relaciones se las denomina *elasticidades*.

Modelo lineal en parámetros

• Modelo **semi-logarítmico**:

$$\log(Y_i) = \beta_1 + \beta_2 X_{2i} + u_i$$

• En este caso, el coeficiente se intrepreta como:

$$\beta = \frac{dY / Y}{dX}$$

• Esto es, cómo cambia porcentualmente Y ante un cambio de una unidad en la variable X. Este tipo de relaciones se las denomina *semi-elasticidades*.

Modelo lineal en parámetros

• Modelo **cuadrático**:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{2i}^2 + u_i$$

• En este caso, el coeficiente se intrepreta como:

$$\frac{dY}{dX} = \beta_2 + 2\beta_3 X_2$$

• Es importante notar que, en este caso, el efecto de X en Y no es constante, sino que depende de en qué valor de X lo evaluemos (Ej: efecto de la edad en los salarios)

Modelo lineal en parámetros

• Modelo con **interacciones**:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{2i} + \beta_4 X_{1i} X_{2i} + u_i$$

• En este caso, el coeficiente se intrepreta como:

$$\frac{dY}{dX_2} = \beta_2 + \beta_4 X_3$$

• En este caso, el efecto de X_2 en Y no es constante, sino que depende de en qué valor de X_3 lo evaluemos (Ej: interacción de la educación con la inteligencia)

Variables explicativas binarias:

• Supongamos que queremos ver el efecto del género en el alguna variable. Para esto, armamos una variable binaria

$$D_i = \begin{cases} 1 & \text{si es hombre} \\ 0 & \text{si es mujer} \end{cases}$$

• Y luego hacemos la regresión

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + u_i$$

• Lo importante cómo es la interpretación del coeficiente que acompaña a la variable binaria

Variables explicativas binarias

• Notemos que para los hombres (D_i=1) el valor esperado de la variable explicada es

$$Y_i^H = \beta_1 + \beta_2 X_i + \beta_3 1 = \beta_1 + \beta_2 X_i + \beta_3$$

• Mientras que para las mujeres (D_i=0) esta relación es

$$Y_i^M = \beta_1 + \beta_2 X_i + \beta_3 0 = \beta_1 + \beta_2 X_i$$

• Luego,

$$\beta_3 = Y_i^H - Y_i^M$$

• De esta manera, el coeficiente que acompaña a la variable binaria se interpreta como cuánto mayor es (en unidades) el valor esperado de la variable explicada para los hombres comparado con las mujeres.

Variables explicativas binarias

• Es importante notar que la comparación es siempre en términos relativos a la categoría base. Esto es porque no podemos incluir dos variables, una que valga 1 cuando sea hombre y 0 si es mujer, y otra que valga 1 cuando es mujer y 0 cuando es hombre, porque tendríamos un problema de multicolinealidad

Variables explicativas binarias

• Supongamos que tenemos más de una categoría, por ejemplo si la variable pregunta que marca de cigarrillos fuma y los encuestados solo responden Marlboro, Camel o Gitans.

• Armemos 2 variables binarias

$$D_{1i} = \begin{cases} 1 & \text{si fuma Marlboro} \\ 0 & \text{si no fuma Marlboro} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{si fuma Camel} \\ 0 & \text{si no fuma Camel} \end{cases}$$

• Recordar: no podemos poner 3 por el problema de multicolinealidad

Variables explicativas binarias

• Luego, estimamos el modelo

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_{1i} + \beta_4 D_{2i} + u_i$$

• Para ver la interpretación de los coeficientes, procedamos como antes

$$Y_i^M = \beta_1 + \beta_2 X_i + \beta_3$$
$$Y_i^C = \beta_1 + \beta_2 X_i + \beta_4$$
$$Y_i^G = \beta_1 + \beta_2 X_i$$

• Notar, que la interpretación es siempre relativa a la categoría base.

Variables explicativas binarias

• Volviendo al caso de diferencias según el sexo, supongamos el siguiente modelo:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i D_i + u_i$$

• Aquí la variable binaria interactúa con la variable *X*. Para ver la interpretación del coeficiente, notemos que

$$\left.\frac{dY}{dX}\right|^{H} = \beta_2 + \beta_3 \quad \text{y} \quad \left.\frac{dY}{dX}\right|^{M} = \beta_2$$

• Luego, el coeficiente se interpreta cómo es el efecto de la variable *X* en *Y* para los hombres en relación a las mujeres

Variables explicativas binarias

• Por último, siguiendo en el caso de diferencias según el sexo, consideremos el siguiente modelo:

$$\log(Y_i) = \beta_1 + \beta_2 X_i + \beta_3 D_i + u_i$$

• En este caso, el coeficiente que acompaña a la variable binaria se interpreta como cuánto mayor es (en porcentaje) el valor esperado de la variable explicada para los hombres comparado con las mujeres.