1.Variables ficticias en el modelo de regresión: ejemplos.

Las variables ficticias recogen los efectos diferenciales que se producen en el comportamiento de los agentes económicos debido a diferentes causas como las siguientes:

- *De tipo temporal:* Para recoger efectos diferentes en función del tiempo en que se producen las observaciones de las variables (por ejemplo, consumo en periodos de guerra o paz).

- *De carácter espacial:* Para tener en cuenta la pertenencia o no de la observación a una determinada zona (por ejemplo, consumo en zonas rurales o urbanas).

- *De tipo cualitativo:* Para recoger los efectos de variables cualitativas como el género, el estado civil, tener o no cargas familiares, nivel de educación, etc. sobre el comportamiento de los agentes económicos en decisiones de consumo, de oferta de trabajo, etc.

- *Otras causas:* Para conocer los efectos que las variables cuantitativas tienen sobre la variable endógena, distinguiendo por submuestras (por ejemplo, la propensión marginal al consumo de individuos de rentas altas o bajas).

2. Interpretación de los efectos de las variables explicativas ficticias: Tipos de modelos.

Para interpretar los efectos de las variables explicativas ficticias en un modelo de regresión se utiliza un ejemplo sencillo. Se supone que tenemos una muestra de individuos ocupados y una característica ocupacional que indica si el individuo es

licenciado o no. A partir de este supuesto se pretende explicar el nivel salarial de los individuos y para ello se plantea la siguiente regresión:

$$Y_i = \hat{a}_1 + \hat{a}_2 X_{2i} + u_i \qquad (1)$$

donde Y_i es el salario individual y X_{2i} es una variable ficticia que toma el valor 1 si el individuo es licenciado y 0 en caso contrario. En (1) β_1 mide el salario esperado de un trabajador no licenciado y β_2 mide la diferencia entre los salarios esperados del trabajador licenciado y no licenciado. Estos efectos de los parámetros se pueden comprobar si se toman esperanzas de la expresión (1). Así:

$$E(Y_i) = \begin{cases} \hat{a}_1 \text{ si } X_{2i} = 0 \\ \hat{a}_1 + \hat{a}_2 \text{ si } X_{2i} = 1 \end{cases} \qquad (2)$$

Existe un test relevante que es contrastar H_0:β_2=0. Si se acepta esta hipótesis, no hay diferencias salariales entre trabajadores licenciados y aquellos que no lo son.

Asimismo, si se considera que en lugar de tener una variable de cualificación con dos valores distintos (licenciado o no), se tiene que hay trabajadores con tres niveles diferentes de cualificación (licenciado, diplomado y no cualificado) que tienen diferente salario. Para contrastarlo se plantea la siguiente ecuación:

$$Y_i = \hat{a}_1 + \hat{a}_2 X_{2i} + \hat{a}_3 X_{3i} + u_i \qquad (3)$$

donde Y_i es el salario individual; X_{2i} es una variable ficticia que toma el valor 1 si el individuo es licenciado, 0 en caso contrario; y X_{3i} toma el valor 1 si el individuo es diplomado, 0 en caso contrario. Tomando valores esperados del salario en la expresión (3):

$$E(Y_i) = \begin{cases} \hat{a}_1 \ \text{si } X_{2i} = 0 \ \text{y } X_{3i} = 0 \\ \hat{a}_1 + \hat{a}_2 \ \text{si } X_{2i} = 1 \ \text{y } X_{3i} = 0 \\ \hat{a}_1 + \hat{a}_3 \ \text{si } X_{2i} = 0 \ \text{y } X_{3i} = 1 \end{cases} \quad (4)$$

donde β_1 mide el salario esperado de un trabajador no cualificado; β_2 mide la diferencia entre el salario esperado de un trabajador licenciado y no cualificado; y β_3 la diferencia entre el salario esperado de un trabajador diplomado y no cualificado. Nuevamente, existen contrastes relevantes[1]. Por ejemplo, si se contrasta H_o:β_2=0 y se acepta la hipótesis, no hay diferencias entre los salarios medios de los trabajadores licenciados y no cualificados. Tampoco habría diferencias salariales entre los trabajadores diplomados y no cualificados si se acepta la hipótesis H_o:β_3=0.

Finalmente, se podría contrastar si hay diferencias entre los salarios medios de los trabajadores licenciados y diplomados a partir de la expresión (3). En este caso se podría plantear un test "F" para la hipótesis nula H_o:β_2=β_3. Sin embargo, operando con las variables ficticias del modelo se puede realizar un contraste más sencillo mediante la distribución "t". Si se escribe el modelo de regresión (3) como:

$$Y_i = g_1 + g_2 X_{2i} + g_3 (X_{2i} + X_{3i}) + u_i \quad (5)$$

[1] Cuando se introducen variables ficticias en un modelo de regresión y el atributo está compuesto de "m" alternativas, se deben incluir m-1 variables ficticias. De lo contrario, se produce un problema de multicolinealidad perfecta conocido como *trampa de las variables ficticias* en el modelo a no ser que se excluya la constante cuando se incluyen "m" variables cualitativas.

estando Y_i , X_{2i} y X_{3i} definidas como antes, y expresando el valor esperado del salario como:

$$E(Y_i) = \begin{cases} \tilde{a}_1 \text{ si } X_{2i} = 0 \text{ y } X_{3i} = 0 \\ \tilde{a}_1 + \tilde{a}_2 + \tilde{a}_3 \text{ si } X_{2i} = 1 \text{ y } X_{3i} = 0 \\ \tilde{a}_1 + \tilde{a}_3 \text{ si } X_{2i} = 0 \text{ y } X_{3i} = 1 \end{cases} \qquad (6)$$

Entonces, el contraste se puede realizar sobre la hipótesis $H_0: \gamma_2 = 0$ mediante un "t" ratio.

Los modelos planteados hasta ahora son muy sencillos, y pueden ser poco realistas porque no incluyen otras variables que influyen sobre los salarios de los trabajadores. Si se tiene información no solo de los salarios y el nivel de cualificación sino también de otras variable como la edad, los años de experiencia, el sector de actividad, etc., la incorporación de esas variables se realizaría sin ninguna dificultad. Así, se puede plantear un modelo como:

$$Y_i = \hat{a}_1 + \hat{a}_2 X_{2i} + \tilde{a} Z_i + u_i \qquad (7)$$

donde X_{2i} indica si el trabajador es cualificado o no y Z_i es el número de años de antigüedad en la empresa del trabajador. La expresión (7) se denomina *modelos de variables ficticias de tipo I* En este modelo la cualificación sólo afecta a la constante u ordenada en el origen. De forma que los salarios medios para los trabajadores se expresarían como:

$$E(Y_i) = \begin{cases} \hat{a}_1 + \hat{a}_2 + \tilde{a}E(Z_i) \text{ si se trata de un trabaja dor cualific.} \\ \hat{a}_1 + \tilde{a}E(Z_i) \text{ si se trata de un trabaja dor no cualific.} \end{cases} \qquad (8)$$

Podría darse el caso que la cualificación tuviera efectos sobre los pagos que por antigüedad tienen los trabajadores. Así, el *modelo de variables ficticias tipo II* recogería estos hechos

$$Y_i = \hat{a}_1 + \tilde{a}Z_i + \boldsymbol{j}(Z_i X_{2i}) + u_i \qquad (9)$$

Y los salarios esperados serían:

$$E(Y_i) = \begin{cases} \hat{a}_1 + (\tilde{a} + \boldsymbol{j})E(Z_i) \text{ si se trata de un trabaja dor cualific.} \\ \hat{a}_1 + \tilde{a}E(Z_i) \text{ si se trata de un trabaja dor no cualific.} \end{cases} \qquad (10)$$

El modelo de variables ficticias tipo III no sólo presenta diferencias en la ordenada en el origen como en el de tipo I o cambios en la pendiente como en el de tipo II. Sino también recoge efectos de la cualificación en los salarios medios, como efectos de interacción con la experiencia del trabajador.

$$Y_i = \hat{a}_1 + \boldsymbol{b}_2 X_{2i} + \tilde{a}Z_i + \boldsymbol{j}(Z_i X_{2i}) + u_i \qquad (11)$$

donde los valores esperados de los salarios medios de los trabajadores serían:

$$E(Y_i) = \begin{cases} \hat{a}_1 + \hat{a}_2 + (\tilde{a} + \boldsymbol{j})E(Z_i) \text{ si se trata de un trabaja dor cualific.} \\ \hat{a}_1 + \tilde{a}E(Z_i) \text{ si se trata de un trabaja dor no cualific.} \end{cases} \qquad (12)$$

Incluso se podrían plantear regresiones separadas para cada submuestra, cualificados y no cualificados, y verificar si existen diferencias. De esta forma se evita la introducción de una variable ficticia que aproxime la característica por niveles de cualificación. Las dos regresiones se pueden expresar de la siguiente manera:

Grupo de cualificados: $Y_{ic} = \boldsymbol{a}_1 + \boldsymbol{a}_2 Z_{ic} + u_{ic}$

Grupo de no cualificados: $Y_{inc} = \ddot{e}_1 + \ddot{e}_2 Z_{inc} + u_{inc} \qquad (12b)$

Haciendo el supuesto de igualdad de varianzas entre los dos grupos, la diferencia entre los coeficientes correspondientes al término independiente de las regresiones (12b) coincide con el coeficiente β_2 de la regresión (8). Además, la

diferencia entre los coeficientes correspondientes a la pendiente es igual a los coeficientes asociados a las interacciones de la variable ficticia cualificación con la variable explicativa número de años de antigüedad en la empresa del trabajador, es decir, $\alpha_2-\lambda_2 = \varphi$ en (9), etc. Estas igualdades seguirán siendo validas si sustituimos los coeficientes por sus correspondientes estimadores. Sin embargo, al separar la muestra total en dos grupos, la estimación de la varianza de las perturbaciones difiere de un grupo a otro, y, por tanto, las desviaciones típicas estimadas de los distintos coeficientes variarán de utilizar la ecuación (9) a realizar su estimación con las ecuaciones (12b). Esto provoca diferencias en los valores de los estadísticos t correspondientes a los coeficientes estimados entre las ecuaciones (9) y (12b). Por tanto, la elección entre (9) y (12b) debe tenerse en cuenta si la principal motivación del estudio es conocer cómo afectan de forma diferente el número de años de antigüedad al caso de los individuos cualificados y no cualificados, o bien simplemente, la cuantía de esta diferencia. En el primer caso se utilizará la estimación por grupos, ecuación (12b), mientras que en el segundo caso se puede utilizar la ecuación (9) para todas las observaciones conjuntamente.

Finalmente, se va considerar la utilidad de las variables ficticias para desestacionalizar una serie temporal. Al estudiar la evolución temporal de cualquier magnitud económica utilizando un conjunto de variables explicativas, es conveniente tener en cuenta las variaciones que se producen como consecuencia del fenómeno de la estacionalidad. La estacionalidad es una variación de la serie de periodicidad inferior a un año.

Los fenómenos estacionales son de carácter cultural o institucional, y no están en principio, relacionados con ningún factor estrictamente económico.

Ejemplo de utilidad de las variables ficticias para el tratamiento de la estacionalidad.

Consideremos por ejemplo el Índice de Producción Industrial en España (IPI). Este indicador sufre una caída espectacular durante el mes de Agosto debido a las vacaciones de verano. También, sufre otra más pequeña en el mes de Diciembre por las fiestas de Navidad.

Si el objetivo es estudiar predicciones para el IPI mediante una serie trimestral, la cartera de pedidos (P) se incluye al ser un factor que anticipa las variaciones del IPI, además de tres variables ficticias d_1,d_2,d_3, donde d_1 toma el valor 1 se la observación t-ésima se produce en el segundo trimestre, 0 en caso contrario; d_2 toma el valor 1 si la observación t-ésima se corresponde al tercer trimestre, 0 en caso contrario; d_3 toma el valor 1 si la observación t-ésima corresponde al cuarto trimestre, 0 en caso contrario.

El modelo sería:

$$IPI_t = \alpha_1 + \alpha_2 P_t + \alpha_3 d_1 + \alpha_4 d_2 + \alpha_5 d_3 + u_t$$

Donde $\alpha_3,\alpha_4,\alpha_5$ miden el efecto estacional diferencial con respecto al primer trimestre, que es la categoría de referencia. Un supuesto implícito en esta forma de cuantificar la estacionalidad es que ésta no varía de un año a otro.

3. Variables endógenas cualitativas y tratamiento: modelos de probabilidad lineal, probit y logit.

En este apartado se plantean tres modelos diferentes para el tratamiento de variables endógenas cualitativas binarias. Si se toma el ejemplo que trata de estudiar la participación o no en el mercado de trabajo de la mujer en función de variables como el número de hijos, el salario del marido, el nivel educativo o, la edad, etc. la variable dependiente tiene naturaleza dicotómica. En otras palabras tiene dos opciones: participar en el mercado de trabajo o no formar parte del

mismo. Pasemos primero a analizar el modelo de probabilidad lineal, más tarde el modelo probit y logit.

3.1 El modelo de probabilidad lineal.

Este modelo se puede presentar de la siguiente manera:

$$Y_i = \hat{a}_1 + \hat{a}_2 X_{2i} + \hat{a}_3 X_{3i} + ... + \hat{a}_{ki} X_{ki} + u_{ii} \qquad (13)$$

donde Y_i toma el valor 1 si se elige la primera opción, y 0 en caso contrario; X_{ji} (j=2,….,k) son variables explicativas y u_i es una perturbación aleatoria que cumple las hipótesis expuestas para el modelo clásico de regresión. Para interpretar el modelo expuesto a través de la expresión (13), se pueden tomar esperanzas y considerar que la variable dependiente toma sólo valores 1 y 0.

$$\begin{aligned} &E(Y_i) = \hat{a}_1 + \hat{a}_2 X_{2i} + \hat{a}_3 X_{3i} + ... + \hat{a}_{ki} X_{ki} = \\ &= 1 \bullet P(Y=1) + 0 \bullet P(Y=0) = P(Y=1) \end{aligned} \qquad (14)$$

Los valores predichos para la variable endógena $\hat{Y}_i$ miden la probabilidad de que el individuo i-ésimo elija la primera opción (denotada por el valor 1), dados los valores de las variables explicativas X_{2i}, X_{3i},…X_{ki} para dicho individuo. La estimación de este modelo por mínimos cuadrados ordinarios (MCO) presenta tres inconvenientes que se exponen a continuación:

En primer lugar, las perturbaciones aleatorias "u_i" no siguen una distribución normal. Sino una distribución binomial. No obstante, la forma de la distribución de u_i no es problema porque para una muestra grande la distribución binomial se aproxima a una normal.

En segundo lugar, el término u_i es heterocedastico. La heterocedasticidad conlleva problemas de eficiencia aunque los estimadores por MCO sean

insesgados y consistentes. Tampoco es un gran inconveniente porque se puede realizar una transformación adecuada para que la perturbación aleatoria sea homocedastica.

En tercer lugar, el mayor inconveniente es que no hay ninguna garantía de que las predicciones que el modelo proporciona de Y estén restringidas al intervalo 0 y 1. Este hecho constituye un grave problema asociado con el modelo de probabilidad lineal.

3.2 El modelo probit.

Debido a los inconvenientes manifestados anteriormente en el modelo de probabilidad lineal, se necesita transformar el modelo original de tal manera que restrinja la predicción de Y a estar dentro del intervalo (0,1). Esto requiere trasladar los valores que pertenecen a una recta real a un intervalo, de manera que mantengan las propiedades de un modelo de regresión. Para ello, es necesario utilizar para $E(Y_i)= P_i$ una función de distribución de probabilidad que se escriba como:

$$P_i = F(\hat{a}_1 + \hat{a}_2 X_{2i} + \hat{a}_3 X_{3i} + ... + \hat{a}_{ki} X_{ki}) \qquad (15)$$

Bajo el supuesto de transformación del modelo utilizando una función de distribución de probabilidad uniforme, se obtiene la versión restringida del modelo de probabilidad lineal. No obstante, entre las muchas alternativas para F(.) en (15), las más comunes son la distribución normal (*modelo probit*) y la logística (*modelo logit*).

Para comprender el funcionamiento del modelo, se supone que existe una variable continua latente (no observada) que es función lineal de las variables explicativas:

$$Y_i^* = \hat{a}_1 + \hat{a}_2 X_{2i} + \hat{a}_3 X_{3i} + ... + \hat{a}_{ki} X_{ki} + u_i \qquad (16)$$

Las observaciones de Y_i^* no están disponibles. Estos datos solo se conocen si las observaciones individuales están en una categoría (valores altos de Y_i^*) o en otra (valores bajos de Y_i^*). De esta forma se puede expresar la probabilidad de observar los valores altos de Y_i^* como:

$$\begin{aligned} P_i &= P(Y_i = 1) = P\left[u_i > -(\hat{a}_1 + \hat{a}_2 X_{2i} + \hat{a}_3 X_{3i} + ... + \hat{a}_{ki} X_{ki})\right] = \\ &= 1 - F\left[-(\hat{a}_1 + \hat{a}_2 X_{2i} + \hat{a}_3 X_{3i} + ... + \hat{a}_{ki} X_{ki})\right] = 1 - F(z_i) \end{aligned} \qquad (17)$$

siendo $z_i = -(\hat{a}_1 + \hat{a}_2 X_{2i} + \hat{a}_3 X_{3i} + ... + \hat{a}_{ki} X_{ki})$.

También, se puede calcular la Pr (Y_i=0) mediante el complementario al suceso anterior F(-z_i)).Además, como u_i está distribuida como una normal, y por tanto, también lo está z_i, las probabilidades en (17) se pueden calcular mediante la expresión:

$$P_i = F(z_i) = \frac{1}{\sqrt{2\delta}} \int_{-\infty}^{-z_i} e^{\frac{-t^2}{z}} dt$$

que se corresponde con la función de distribución de la normal estándar.

3.3El modelo logit.

Si se supone que la distribución de F(.) en (15) es la logística, tenemos el modelo logit. La expresión de la función logística es:

$$P_i = F(z_i) = \frac{1}{1 + e^{-z_i}} \qquad (18)$$

siendo “e” la base del logaritmo natural. En realidad, el modelo logit puede estimarse mediante el procedimiento de MCO. De forma que:

$$e^{-z_i} = \frac{1 - P_i}{P_i}$$

y tomando logaritmos naturales queda:

$$\mathrm{Ln}\left(\frac{P_i}{1 - P_i}\right) = \acute{a} + \hat{a}_2 X_{2i} + \ldots + \hat{a}_k X_{ki} \qquad (19)$$

Si se dispone de datos apropiados, es decir de frecuencias para cada individuo, el modelo expresado en (19) se estima por MCO sin dificultad. Sin embargo, la estimación del modelo logit y probit se realiza normalmente por el procedimiento máximo verosímil.