CAPÍTULO III: METODLOGÍA

3.1. Introducción

Una vez enunciada la teoría y la evidencia empírica existente sobre la migración de retorno resulta conveniente pasar a la metodología que se ocupa en este trabajo para probar las hipótesis postuladas al principio.

Para evaluar la primer hipótesis basta recurrir a la estadística descriptiva de los datos, para analizar las características de los migrantes. Esto se realiza en el siguiente capítulo. Para la segunda hipótesis, se llevan a cabo dos metodologías.. La primera consiste de un modelo econométrico probit, y la segunda en el método de la variable instrumental.

El capítulo se organiza de la siguiente manera. En las próximas dos secciones se especifican en detalle las metodologías a ser utilizadas en este trabajo. En la tercera sección se hace referencia a las variables que se ocupan en los modelos. Y en el último apartado se enuncia el posible problema de autoselección.

3.2. Metodología 1: Modelo Probit

3.2.1. Justificación

El objetivo de este trabajo es el de encontrar que tan probable es para un remigrante el convertirse en autoempleado tomando en cuenta las características que tiene. El modelo probit permite llevar a cabo este tipo de estudios dado que en éste se cumple la existencia de una variable latente subyacente para la cual se observa una evidencia dicotómica. En este trabajo, el modelo probit postula como variable observable si la persona tiene o no un

negocio. Con esta información subsecuentemente el modelo reproduce una variable latente, que se define como la propensión que tienen un individuo de abrir un negocio.

3.2.2. El Modelo Probit

Se observa el siguiente modelo $P(y=1/x)=G(\beta_0+\beta_1x_1+...+\beta_kx_k)=G(\beta_0+\beta X)$, en donde *G* es una función que adopta valores entre cero y uno para todos los números reales z. En el modelo probit, *G* representa la función de distribución acumulativa normal estandarizada dada por:

$$F(Z_i)=\int_{-\infty}^{Z_i/\sigma}\frac{1}{\sqrt{2\pi}}\exp[-\frac{t^2}{2}]dt$$

Debido a que el modelo probit es un modelo de variable dependiente limitada, la estimación de los parámetros se hace a través del método de máxima verosimilitud. Este método sugiere que se elijan como estimados los valores de los parámetros que maximicen el logaritmo de la función de verosimilitud (Maddala, 1997). La función logarítmica de verosimilitud para la observación *i* está dada por:

$$\lambda_i(\beta)=y_i\log(G(x_i\beta))+(1-y_i)\log(1-G(x_i\beta))$$

El logaritmo de la función de verosimilitud para una muestra de tamaño *n* se define entonces como: $£=\sum_{i=1}^{n}\lambda_i(\beta)$.

El estimador de máxima verosimilitud de β, denotado por $\hat{\beta}$ maximiza este logaritmo de verosimilitud (Wooldridge, 2003). Las propiedades de los estimadores de

máxima verosimilitud del modelo son consistentes, asintóticamente normales, y asintóticamente eficientes.

A fin de conocer los efectos de los cambios en las variables explicativas sobre las probabilidades de que cualquier observación pertenezca a uno de los dos grupos (y=0, y=1), se emplea una derivada parcial denotada como:

$$\frac{\partial p(x)}{\partial x_j} = g(\beta_0 + x\beta)\beta_j donde, g(z) \equiv \frac{\partial G}{\partial z}(z)$$

El término g(z) corresponde a una función de densidad de probabilidad. Dado que en el modelo probit G(.) es una función de distribución acumulativa estrictamente positiva, g(z)>0 para toda z, el signo del efecto parcial es el mismo que el de β_j.

Para probar la significancia de cada uno de los coeficientes estimados se lleva a cabo la prueba de hipótesis $H_0 : \beta_j = 0$, con un t estadístico $\hat{\beta}_j \Big/ (se)\hat{\beta}$. Para probar la significancia de variables conjuntamente, existen diferentes estadísticos, como el estadístico Wald y el estadístico de la razón de verosimilitud entre otros. En estos dos casos se emplea una distribución chi-cuadrada (Wooldridge,2003).

3.2.3. El Modelo Econométrico

En la revisión bibliográfica dos trabajos realizados por McCormick y Wahba (2001) para el caso de Egipto y Mesnard (1999) para el caso de Tunes utilizan un modelo probit para explicar la decisión de ocupación de los remigrantes. Tomando sus modelos como referencia, se define en este trabajo el modelo econométrico como:

$$y_i^* = \beta_0' X + \beta_1 s_i + \beta_2 m_i + u_i$$

donde *X* es un vector que comprende las variables de control de los individuos, *s* representa los ahorros acumulados durante la permanencia en los Estados Unidos, *m* son los meses de permanencia en el exterior y *u* es el término de error normalmente distribuido con media cero y varianza uno. Dado que no se observa *y**, sólo cuando el individuo se convierte en autoempleado y abre un negocio

$y_i=1$ *si* $y^*>0$ *(si el individuo se convierte en autoempleado)*

$y_i=0$ *si* $y^* \leq 0$ *(si el individuo no se convierte en autoempleado)*

3.3. Metodología 2: El Método de la Variable Instrumental

3.3.1. Justificación

La necesidad de recurrir a esta segunda metodología surge a partir de un posible problema de endogeneidad de la variable explicativa de ahorro en el modelo econométrico postulado previamente. Ocurre el problema de endogeneidad de los ahorros en la ecuación de la decisión de ocupación, si los ahorros de los migrantes son determinados anteriormente durante su permanencia en el exterior.

La evidencia empírica sustenta la idea de endogeneidad del ahorro. Ilahi (1999) señala que durante la estancia en el exterior aquellos migrantes con interés de convertirse en autoempleados a su retorno, ahorran más que los demás migrantes. Mesnard(1999) por su parte muestra que a su retorno, el 87% de los proyectos emprendidos por remigrantes son financiados con sus propios ahorros.

En este trabajo, siguiendo la literatura empírica, se asume la existencia de endogeneidad en las variables explicativas de ahorro del modelo. Por lo tanto, se recurre al método de variables instrumentales para solucionar el problema. A continuación se explica en que consiste este método.

3.3.2. Planteamiento del Problema

Suponiendo el siguiente modelo:

$$y = \beta_0 + \beta_1 x + u$$

donde *y* es la variable dependiente y *x* la variable explicativa, el problema de endogeneidad ocurre cuando la variable *x* se encuentra correlacionada con el término de error *u*. Esto provoca la violación de uno de los supuestos detrás del método de mínimos cuadradas: Cov(x,u)=0. Si no se corrige este problema, se corre el peligro de obtener estimadores sesgados.

3.3.3. El Método de la Variable Instrumental

Para solucionar el problema, se requiere de una nueva variable que cumpla con las siguientes dos restricciones:

$$\text{Cov}(z,u)=0 \qquad (1)$$

$$\text{Cov}(z,x)\neq 0. \qquad (2)$$

donde *z* es la variable instrumental en este caso. Maddala (1997) define la variable instrumental como aquella que no se correlaciona con el término de error, pero sí con la variable explicativa de la ecuación.

En este método se prueba la correlación entre la variable explicativa endógena y la variable instrumental a través de una regresión simple entre estas variables.

$$x = \pi_0 + \pi_1 z + v, \qquad (3)$$

donde $\pi_1 = Cov(z,x)/Var(z)$. Bajo la prueba de hipótesis:

$$H_0 : \pi_1 = 0$$
$$H_a : \pi_1 \neq 0$$

se cumple Cov(z,x)≠0 si se rechaza la hipótesis nula a un nivel de significancia bastante pequeño (entre 1 y 5%).

Si se cumplen los supuestos (1) y (2), se define el parámetro β y el estimador de la variable instrumental como:

$$\beta_1 = \frac{Cov(z,y)}{Cov(z,x)}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

El estimador de la variable instrumental tiene una distribución normal en muestras relativamente grandes. En muestras pequeñas, el estimador puede estar un poco sesgado.

Por otra parte, bajo este modelo el supuesto de homoscedasticidad de las varianzas condicionales de *u* se cumple con respecto a la variable instrumental, no con respecto a la variable independiente endógena. Esto es: $E(u^2/z) = \sigma^2 = Var(u)$. Si la correlación entre la variable endógena y la variable instrumental es pequeña, la varianza de la variable

instrumental puede ser mucho mayor a la varianza del método de mínimos cuadrados. Si por el contrario, existe una alta correlación entre esas variables, la varianza tiende a disminuir.

3.3.4. Evidencia Empírica

3.3.4.1. Caso 1: Tunes

Mesnard (1999) utiliza un modelo probit usando el método de la variable instrumental para evitar problemas de endogeneidad en la variable de ahorro. Se plantea el modelo de la siguiente forma:

$$y_{1i}^{*} = X_{1i}\beta + z_{2i}\gamma + u_{1i}$$
$$z_{2i} = X_{i}\delta + v_{2i}$$

donde $X_i = (X_{1i,} X_{2i})$ es un vector de variables exógenas y z es la acumulación de ahorro. La endogeneidad de esta variable es estimada a través de la siguiente ecuación:

$$y_{1i}^{*} = X_{1i}\beta + z_{2i}\gamma + \hat{v}_{2i}\,\mu + \eta_{1i}$$

en donde, $\hat{v}_{2i} = z_{2i} - X_i\hat{\delta}$ es el término de error de la regresión de mínimos cuadrados de z_{2i} con X_i y $\eta_{1i} = u_{1i} - \hat{v}_{2i}\mu$. Se llega a la conclusión de que $\mu=0$ por lo que la hipótesis de exogeneidad de los ahorros no se rechaza.

3.3.4.2. Caso 2: Pakistán

Ilahi (1999) no conduce prueba de endogeneidad de la variable de ahorro. Lo toma como dado y utiliza el método de mínimos cuadrados ordinarios con un modelo recursivo para solucionar el problema de endogeneidad. En su modelo, se plantean dos ecuaciones:

$$S_i = \psi^1(X_1) + \varepsilon_{1i}$$
$$J_i = \psi^2(X_2, S_i) + \varepsilon_{2i}$$

donde X representa vectores de las variables exógenas del modelo. S son los ahorros hechos en el exterior y J es una variable index. El método que se sigue en este artículo es el de estimar en primer lugar la ecuación de ahorros por mínimos cuadrados. Y de esa regresión se obtiene los valores ajustados $(\hat{S}_i)$. En segundo lugar, se estima la ecuación de decisión de ocupación, en la cual la variable $(\hat{S}_i)$ junto con las otras variables exógenas ocupan el lugar derecho de la ecuación.

3.4. Variables del Modelo

En este trabajo se realizan dos modelos econométricos. En el primero de ellos se analizan únicamente los negocios con establecimientos fijos; en el segundo se agrupan los negocios agrícolas. El objetivo que se persigue con esta separacion es el de medir el impacto que la acumulación de capital humano y financiero tienen en la creación de un negocio agrícola como no agrícola. Se esperaría encontrar que los negocios con establecimientos fijos requieren de una mayor acumulación de capital.

Los dos modelos tienen como variable dependiente, una variable categórica *negfijo* y *negagricultura* respectivamente, las cuales adquieren el valor de uno y cero dependiendo si el remigrante decide invertir en un negocio. Ambas variables se limitan a tomar en cuenta a aquellos negocios que son abiertos ya sea en el mismo año o después, del primer viaje migratorio del jefe de familia.

En la base de datos, la variable negocios con establecimientos fijos se restringe a la inversión en una tienda, un restaurant, un tallar o una fábrica. La variable negocios agrícolas agrupa a todos los negocios relacionados con la agricultura y la crianza de ganado.

Ambos modelos comprenden las mismas variables independientes. Según la especificación de la ecuación del modelo hecha anteriormente, el vector *X* abarca las variables de control de los individuos. A fin de facilitar el entendimiento de las mismas, el vector *X* se subdivide en cuatro categorías, las cuales son: características personales del individuo, características de la migración del individuo, rasgos sobre el tipo de negocio implementado y por último, variables relacionadas con el lugar y el año de encuesta. A continuación se hace una breve descripción de estos grupos de variables incorporados en el modelo, incluyendo la explicación de las variables *s* y *m* del modelo.

3.4.1. Características Personales

Bajo esta división se especifican todas las variables que describen a grandes rasgos a los jefes de familia. Estas variables son: *edad, estado civil, si es padre, número de hijos* y *educación.*

La variable *edad* es una variable numérica que describe la edad del jefe de familia en el momento de la encuesta. Esta variable está expresada en años. La variable *estado civil* está especificada en el modelo a través de seis variables categóricas que son: *soltero, casado, unión libre, viudo, divorciado* y *separado*. La variable categórica *casado* es omitida del modelo puesto que es la variable de referencia

Con relación al *número de hijos*, se agregan al modelo dos variables. La primera, *es papá* es una variable categórica que toma el valor de uno y cero dependiendo si el jefe de familia tiene hijos o no. Y la segunda, *hijos* es una variable numérica del número de hijos que tiene el jefe de familia viviendo en la misma casa que él.

También se añaden al modelo diez variables categóricas para distinguir entre el nivel de estudios del jefe de familia. Las variables son: *sin escolaridad, primaria completa e incompleta, secundaria completa e incompleta, preparatoria completa e incompleta, universidad completa e incompleta y posgrado.*

Se entiende por completa cuando los individuos alcanzan los seis, nueve, doce y diecisiete años de estudio respectivamente. De lo contrario se coloca como incompleto el nivel de estudio. En el trabajo se asume que los años que comprende el estudio superior son cinco. La variable *primaria incompleta* queda excluida del modelo porque es la variable de referencia.

3.4.2. Características Migratorias

En cuanto a las características de la migración, las variables de este apartado también explican las características de lo jefes de familia, pero durante su experiencia migratoria en los Estados Unidos. Las variables de este apartado son: *numero de viajes, permanencia en los Estados Unidos del primer y último viaje, edad de retorno del primer y último viaje, ahorros acumulados en el país huésped, ahorros traídos a México, cantidad de remesas, conocimiento del inglés* y *ocupación en los Estados Unidos durante el último viaje*.

La variable numérica *numero de viajes* expresa el número de viajes que el individuo ha realizado a los Estados Unidos hasta el momento de la encuesta. Dado que la base analiza únicamente a los jefes de familia con experiencia migratoria, el valor inicial de esta variable es uno, lo cual se traduce a un solo viaje al otro país. En cuanto a la *edad de retorno del primer y último viaje*, se crean dos variables numéricas *edadretornopv y edadretornouv.* Ambas variables están medidas en años.

Por otra parte, la variable *m* incorporada en el modelo econométrico está denotado por las variables *duracionpv* y *duracionuv,* que expresan la duración de la permanencia en el exterior para el primer y último viaje, respectivamente. Con estas variables se busca medir la relación que existe entre el tiempo de residencia en el extranjero con la decisión de ser autoempleado al retorno de la migración. Estas variables están medidas en meses.

La acumulación de capital financiero en el exterior *s*, esta descrita por tres variables. La primera de ellas, *ahorroseu,* mide la cantidad de dólares ahorrados mensualmente durante la estancia del individuo en los Estados Unidos. La segunda de ellas, *ahorrosmx,* calcula el monto que los migrantes traen a México a su retorno. Y la tercera *remesas,* determina el valor mensual de dinero que es enviado a México. Las tres variables están expresadas en dólares. Y no existe una fuerte correlación entre las tres variables.

Resulta peculiar no encontrar correlación entre estas tres variables. La justificación de tal fenómeno se sustenta en los siguientes argumentos. En primer lugar, los ahorros que los individuos acumulan en los Estados Unidos pueden no ser destinados en su totalidad a remesas o a guardarlos y traerlos al retorno. Puede ser que parte de lo ahorrado

permanezca guardado con familiares o amigos en los Estados Unidos, para el siguiente viaje migratorio que realice el individuo. Esto suponiendo que existe una gran circularidad migratoria por parte del individuo.

En segundo lugar, generalmente el sueldo percibido por los migrantes indocumentados no varía mucho, mientras que la cantidad de remesas enviadas si. Se ha observado que dependiendo de las necesidades del hogar mexicano, se determina la cantidad de remesas enviadas. Por lo tanto, personas ahorrando la misma cantidad de dinero en los Estados Unidos pueden diferir en la cantidad de envíos de remesas..

En tercer lugar, es difícil determinar el patrón de ahorros de los migrantes, porque puede darse el caso de que estos ganen muy poco más sin embargo manden mucho dinero a sus hogares. Disminuyendo los costos de manutención, como lo son el comer poco, o vivir en una casa o cuarto con muchos migrantes, se puede lograr tal comportamiento.

Por otra parte, para poder cuantificar la acumulación de capital humano durante la experiencia migratoria, se introducen al modelo tanto variables relacionadas con la ocupación del individuo en los Estados Unidos, como variables que especifican el nivel de conocimiento del inglés.

Para el caso de la ocupación en el exterior, se sigue la misma clasificación determinada por el apéndice D de la encuesta *MMP93*, en la cual se dividen las ocupaciones en diecinueve opciones. La variable de referencia utilizada es *trabajador en agricultura.* Ver Anexo.

La clasificación de la variable de idiomas se realiza en base a su nivel de comprensión y habla. Las cinco variables categóricas que se generan son *nohablanoentiende, nohablaentiendealgo, nohablaentiendemucho, hablaentiendealgo* y *hablaentiendemucho*. En este caso, la variable de referencia es *nohablonoentiendo.*

3.4.3. Características del Negocio

Las variables de este grupo tienen por finalidad indagar sobre las características del negocios en el que se decide invertir. Las variables de esta división son: *tipo de tierra que se tiene,* y *tenencia que se tiene sobre la tierra.* En el modelo de negocio fijo con establecimiento no se incorporan estas variables, puesto que resultan innecesarias en el modelo. Éstas sólo se utilizan en el modelo de negocio agrícola.

Para definir la variable *tipo de tierra que se tiene*, se crean seis variables categóricas que son: *de irrigación, tierra húmeda, tierra seca, huerto, pastizal* y *otros*. La variable *otros* queda fuera del modelo por ser la variable de referencia.

Y en cuanto a la *tenencia de la tierra*, también se generan variables categóricas *ejido, privado, comunal* y *alquilada.* La variable *alquilada* se toma como variable de referencia. Cabe mencionar, que en el caso de las variables de negocio, el jefe de familia no es la unidad de análisis, sino que lo es el hogar.

3.4.4. Características de Lugar

La última división que se hace es la de lugar. Este grupo comprende las variables *lugar de residencia en los Estados Unidos, lugar de encuesta, y año de encuesta.* El objetivo que se

busca con estas variables es el de identificar el efecto que el lugar y el tiempo tienen sobre la decisión de invertir en un negocio.

Para identificar el *lugar de encuesta*, se crean cuatro variables categóricas *areametropolitana, urbanopequeño, pueblo* y *rancho*. Las primeras dos variables representan áreas urbanas, mientras que las otras dos, áreas rurales. La variable de referencia en este caso es *rancho*. Para denotar los *años de encuesta*, también se forman variables categóricas para todos los años en los que se llevó a cabo la encuesta, comenzando en *1982* hasta llegar al *2002*. La variable de referencia es *1991*.

Por último, se eligen los diez estados más visitados por los migrantes mexicanos en su último viaje a los Estados Unidos, y de esos estados se generan variables categóricas. Los estados analizados son: *California, Arizona, Illinois, Nueva York, Colorado, Pennsylvania, Idaho, Texas, Florida* y *Nevada.* Los *otros estados* excluidos del análisis representan el grupo de referencia.

3.5. El Problema de Autoselección

En la base de datos utilizada en este trabajo se observa una pérdida importante de observaciones. La causa de esta pérdida de información se le atribuye a dos factores: en primer lugar, en muchos casos los encuestados no contestan a las preguntas dado que la pregunta no es aplicable a su caso, por lo que en su respuesta se contesta como no aplica; en segundo lugar, la gente puede no contestar porque no saben con exactitud la respuesta a

la pregunta, o no quieren revelar la información, por lo que en su respuesta se contesta como desconocido.

En este trabajo, tanto la respuesta no aplica, como la respuesta desconocido son tratados como datos faltantes. A consecuencia de eso, las regresiones generan pérdidas en el número de observaciones.

A fin de determinar si la pérdida de observaciones es sustancial en la interpretación de los resultados, y con el objetivo de asegurar la validez de los mismos, se verifica la existencia de un problema de autoselección mediante dos tablas estadísticas. En cada tabla se muestra una descripción estadística de las variables del modelo, incorporando en la primera tabla las 5234 observaciones originales, y en la segunda, las 3563 observaciones trabajadas en los modelos. De obtener resultados muy distintos entre las dos tablas, se estaría incurriendo en un problema de autoselección.