Serie Apuntes de Clase ΩΒΓ

N° 05
Setiembre de 2014

ECONOMETRIA DE DATOS DE PANEL.

Rafael Bustamante Romani

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú, DECANA DE AMÉRICA

FACULTAD DE CIENCIAS ECONÓMICAS

La **Serie Apuntes de Clase Omega Beta Gamma** tiene por objetivo difundir los materiales de enseñanza generados por los docentes que tienen a su cargo el desarrollo de las asignaturas que forman parte de los Planes de Estudios de las Escuelas Académico-Profesionales de la Facultad de Ciencias Económicas de la Universidad Nacional Mayor de San Marcos. Estos documentos buscan proporcionar a los estudiantes la explicación de algunos temas específicos que son abordados en su formación universitaria.

Encargados de la serie:

Bustamante Romaní, Rafael.
rbustamanter@unmsm.edu.pe

Cisneros García, Juan Manuel.
jcisnerosg@unmsm.edu.pe

Facultad de Ciencias Económicas.
Universidad Nacional Mayor de San Marcos.
Calle Germán Amézaga N° 375.
Ciudad Universitaria, Lima 1. Perú.

La **Serie Apuntes de Clase ΩBΓ** es promovida y desarrollada por un colectivo de docentes del Departamento de Economía de la Universidad Nacional Mayor de San Marcos.
El contenido de cada publicación es íntegramente responsabilidad de cada autor, no representa necesariamente los puntos de vista de los integrantes del colectivo, ni de la Universidad.

Econometría de Datos de Panel.*

Rafael Bustamante Romaní◊

Resumen

El aumento de bases de datos, junto con el progreso en las técnicas econométricas, ha facilitado el perfeccionamiento de estudios cada vez más sofisticados de los fenómenos económicos, permitiendo asesorar más acertadamente a los responsables de la elaboración de las políticas públicas y a los hombres de negocios. Sin embargo, estas herramientas se han tornado cada vez más complejas, demandando un alto grado de conocimiento teórico y práctico para poder implementarlas. La metodología de Datos de Panel es una de las más usadas en los últimos tiempos en el ámbito de la economía, las finanzas y los negocios. Su riqueza radica en que permite trabajar simultáneamente varios periodos de tiempo y los efectos individuales, y a su vez, tratar el problema de la endogeneidad. A pesar de las ventajas de esta técnica, existen diversos obstáculos para su implementación, tanto metodológicos como operativos. Esta guía intenta ayudar a los alumnos, investigadores y profesionales que buscan llevar a cabo estudios utilizando Datos de Panel, ofreciendo una pauta para manejar y analizar datos, en forma conjunta con revisar sus fundamentos.

Palabras Claves: Métodos econométricos: modelos de ecuaciones múltiples; Datos de Panel.

Clasificación JEL: C3, C33

◊ Estudios concluidos de Doctorado en Economía con mención en los Recursos Naturales (c), Universidad Nacional Autónoma de México. MBA Gerencial, CENTRUM Pontificia Universidad Católica del Perú. Maestría en Economía con mención en Finanzas, Universidad Nacional Mayor de San Marcos. B. Sc. Economía, UNMSM. Profesor Auxiliar del Departamento de Economía de la UNMSM. Investigador asociado al Instituto de Investigaciones FCE - UNMSM. Contacto: rbustamanter@unmsm.edu.pe.

El autor agradece la colaboración en la elaboración del presente documento a Oscar Díaz Barzola, alumnos de la facultad de ciencias Económicas. Contacto oscardiaz96@outlook.com

I. INTRODUCCIÓN

Los modelos que se utilizan en el análisis económico o de cualquier otra índole, se pueden clasificar desde dos puntos de vista: i) según los datos utilizados y ii) según relaciones supuestas entre las variables. El objetivo de este documento, es analizar el primero de los dos criterios.

En el análisis de la información (económica, social, empresarial, comercial, etc.) pueden existir diferentes dimensiones sobre las cuales interesa obtener conclusiones derivadas de la estimación de modelos que traten de extraer relaciones de causalidad o de comportamiento entre diferentes tipos de variables, a partir de los datos disponibles (de Arce & Mahía, 2007).

Una de estas dimensiones la constituye el análisis de series de tiempo, la cual incorpora información de variables y/o unidades individuales de estudio durante un período determinado de tiempo (dimensión temporal). En este caso, cada período de tiempo constituye el elemento poblacional y/o muestral. Por su parte, existe otra dimensión que no incorpora el aspecto temporal sino que más bien representa el análisis de la información para las unidades individuales de estudio, en un momento determinado del tiempo (dimensión estructural). En este tipo de análisis, el cual se denomina de corte transversal, el elemento o unidad maestral no lo constituye el tiempo sino las unidades de análisis.

Ejemplos de este tipo de análisis pueden ser la cantidad demandada de alimentos por una muestra de familias durante un período de tiempo o la cantidad producida de televisores por una serie de empresas en el mismo lapso. En ambos casos los elementos muéstrales serían la familia y la empresa.

Los dos tipos de análisis de la información permiten extraer conclusiones relevantes de acuerdo con los intereses del investigador. Un sencillo ejemplo puede ilustrar mejor la diferencia entre cada uno de los enfoques: supóngase que se quiere modelar la rentabilidad de las firmas que pertenezcan a una determinada industria. Un análisis de regresión basado en datos de corte transversal para un año en particular podría incluir un conjunto de variables explicativas tales como la calidad de la administración, el monto del capital físico, el empleo de mano de obra y el nivel de apalancamiento financiero, por ejemplo. Con el tipo de información incluida en este modelo de corte transversal se podría estar tomando en cuenta cualquier tipo de economía de escala de la que las firmas podrían beneficiarse. Sin embargo, este modelo no podría identificar o tomar en cuenta como variable explicativa de la rentabilidad, cualquier incremento en el rendimiento que pueda ocurrir con el transcurso del tiempo como consecuencia de mejoras tecnológicas que hayan sido incorporadas en la industria (de Arce & Mahía, 2007).

El principal objetivo de aplicar y estudiar los datos en panel, es capturar la heterogeneidad no observable, ya sea entre agentes económicos o de estudio así como también en el tiempo, dado que esta heterogeneidad no se puede detectar ni con estudios de series temporales ni tampoco con los de corte transversal.

Esta técnica permite realizar un análisis más dinámico al incorporar la dimensión temporal de los datos, lo que enriquece el estudio, particularmente en períodos de grandes cambios. Esta modalidad de analizar la información en un modelo de panel es muy usual en estudios de naturaleza microeconómica.

La aplicación de esta metodología permite analizar dos aspectos de suma importancia cuando se trabaja con este tipo de información y que forman parte de la heterogeneidad no observable: i) los efectos individuales específicos y ii) los efectos temporales (M., 2000).

En lo que se refiere a los efectos individuales específicos, se dice que estos son aquellos que afectan de manera desigual a cada uno de los agentes de estudio contenidos en la muestra (individuos, empresas, bancos) los cuales son invariables en el tiempo y que afectan de manera directa las decisiones que tomen dichas unidades. Usualmente se identifica este tipo de efectos con cuestiones de capacidad empresarial, eficiencia operativa, capitalización de la experiencia, acceso a la tecnología, etc.

Los efectos temporales serían aquellos que afectan por igual a todas las unidades individuales del estudio pero que no varían en el tiempo. Este tipo de efectos pueden asociarse, por ejemplo, a los choques macroeconómicos que pueden afectar por igual a todas las empresas o unidades de estudio.

Los datos longitudinales o paneles siguen a los elementos de la muestra a lo largo del tiempo, recogiendo observaciones consecutivas de los mismos elementos. Una encuesta longitudinal hace posible analizar cómo cambia el valor de determinadas variables a lo largo del tiempo para elementos concretos de la muestra. El número de individuos (N) es generalmente grande y el número de períodos (T) es generalmente reducido. Una encuesta longitudinal hace posible analizar cómo cambia el valor de determinadas variables a lo largo del tiempo para elementos concretos de la muestra.

II. DIVERSAS DEFINICIONES

- **Panel Data** es mezclar información de corte transversal e información temporal. Como en el corte transversal, se recoge información de individuos y se observa cada individuo, como en el análisis de series de tiempo, a través del tiempo. Esto permite estudiar los efectos dinámicos y de comportamiento individual de los problemas.

- Son observaciones repetidas sobre el mismo conjunto de unidades de sección cruzada o dicho de otra forma se tiene el mismo número de observaciones en cada unidad de sección cruzada es decir es una mezcla de ambas en la cual se recoge información entre individuos y se observa cada individuo como el análisis de series de tiempo, a través del tiempo.

- En los paneles microeconómicos, el investigador está interesado en analizar como varía el comportamiento de los agentes económicos individuales frente a cuestiones como sus hábitos de consumo, su situación laboral, su nivel de estudios, etc. Estas son decisiones que dependerán de una lista de características socioeconómicas que el analista debe especificar como variables explicativas del modelo. Sin embargo no todos los agentes toman sus decisiones de igual modo: diferentes agentes, incluso si comparten las mismas características observables, toman decisiones distintas. Ello obliga a contemplar la existencia de efectos no observables, específicos de cada agente encuestado, generalmente constantes en el tiempo, que inciden sobre el modo en que este toma sus decisiones. Si estos efectos latentes existen y no se recogen explícitamente en el modelo, se producirá un problema de variables omitidas: los coeficientes estimados de las variables explicativas incluidas estarán sesgados, por recoger parcialmente los efectos individuales no observables (Greene, 1999)

Para entender mejor esta metodología veamos algunos aspectos matriciales.

$$Y_i = \begin{bmatrix} \overline{Y_{i1}} \\ Y_{i2} \\ . \\ . \\ . \\ Y_{iT} \end{bmatrix}_{TX1} \quad Para\ todo\ t\ = 1,\ 2,3,....,\ T \tag{1.}$$

$$X_i = \begin{bmatrix} X_{i1}{}^{1} & X_{i1}{}^{2} & . & . & . & X_{i1}{}^{K-1} & X_{i1}{}^{k} \\ X_{i2}{}^{1} & X_{i2}{}^{2} & . & . & . & X_{i2}{}^{K-1} & X_{i2}{}^{K} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ X_{iT-1}{}^{1} & X_{iT-1}{}^{2} & . & . & . & X_{iT-1}{}^{K-1} & X_{iT-1}{}^{K} \\ X_{iT}{}^{1} & X_{iT}{}^{2} & . & . & . & X_{iT}{}^{K-1} & X_{iT}{}^{K} \end{bmatrix}_{TxK} \tag{2.}$$

Además los errores del modelo y la variable explicativa se expresan:

$$\varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ . \\ . \\ . \\ \varepsilon_{iT-1} \\ \varepsilon_{iT} \end{bmatrix}_{TX1} \quad Ademas: \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_{N-1} \\ \varepsilon_N \end{bmatrix}_{NTX1} \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_{N-1} \\ Y_N \end{bmatrix}_{NTX1} \qquad 3.$$

$$X = \begin{bmatrix} X_1 \\ X_2 \\ . \\ . \\ . \\ X_{N-1} \\ X_N \end{bmatrix}_{NTXK} \quad \forall t = 1, 2, 3, T \quad i = 1, 2, 3, \ldots.., N \qquad 4.$$

El modelo totalmente apilado es:

$$\mathrm{Y} = \mathrm{X}\beta + \varepsilon \qquad \textbf{5.}$$

X_{jit} : Es el valor Jth de una variable explicativa i. Para todo $t = 0\ 1, 2,3,\ldots.$ T.

Si existen K variables explicativas el vector de variables explicativas se puede denotar como:

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_{K-1} \\ \beta_K \end{bmatrix} \qquad \mathbf{6.}$$

En base a lo anotado podemos afirmar que metodología de datos de panel lo que hace es utilizar procedimientos adecuados para el manejo de las observaciones con una dimensión de sección cruzada grande, con el objeto de estimar modelos econométricos que incluyan entre las variables explicativas los efectos individuales no observables.

El disponer de un número reducido, T, de observaciones de cada uno de los N individuos de la muestra, podría pensarse en estimar un modelo econométrico con cada una de las T secciones cruzadas para luego comparar la evolución de los coeficientes del modelo a lo largo del tiempo.

Las ventajas de modelos econométricos con información en panel, son las siguientes (Greene, 1999):

- Se dispone de un gran número de datos (a través de individuos y a través del tiempo). Por esta razón aumentan los grados de libertad y, al utilizar las diferencias individuales en los valores de las variables explicativas, se reduce la colinealidad entre las variables explicativas, mejorando de esta forma la eficiencia de los estimadores.

- Evita los sesgos de agregación con datos macroeconómicos.

- En general, es posible obtener estimaciones consistentes para $N \to \infty$ y T fijo. No obstante, dada la creciente existencia de bases de datos longitudinales con períodos muéstrales prolongados, existen trabajos recientes en que se consideran propiedades asintóticas para $N \to \infty$ y $T \to \infty$.

- La disponibilidad de datos longitudinales permite a los investigadores analizar una variedad de importantes interrogantes económicas, que no se pueden analizar utilizando solo información de corte transversal o sólo información de series de tiempo.

- Permite construir y testear modelos de comportamiento más sofisticados que los modelos econométricos estándar de series de tiempo o de corte transversal.

- Proporciona un método para resolver o reducir la magnitud de un problema econométrico clave que siempre surge en los trabajos empíricos: siempre se señala que la verdadera razón de porque se encuentra (o no se encuentran) ciertos efectos es producto de la omisión de variables- debido a problemas de medición o porque ciertas variables no son observadas – que están correlacionadas con las variables explicativas.

- Permite estudiar de una mejor manera la dinámica de los procesos de ajuste. Esto es fundamentalmente cierto en estudios sobre el grado de duración y permanencia de ciertos niveles de condición económica (desempleo, pobreza, riqueza).

- Permite elaborar y probar modelos relativamente complejos de comportamiento en comparación con los análisis de series de tiempo y de corte transversal. Un ejemplo claro de este tipo de modelos, son los que se refieren a los que tratan de medir niveles de eficiencia técnica por parte de unidades económicas individuales (empresas, bancos, etc.) (Beltrán Barco, 2003).

- Permite al investigador mucha más flexibilidad para modelizar las diferencias de comportamientos entre los individuos. Tal y como se mencionó anteriormente, la técnica permite capturar la heterogeneidad no observable ya sea entre unidades individuales de estudio como en el tiempo. Con base en lo anterior, la técnica permite aplicar una serie de pruebas de hipótesis para confirmar o rechazar dicha heterogeneidad y cómo capturarla.

Entre las desventajas se cuentan:

- **Sesgo de heterogeneidad:**

Muchos paneles de datos provienen de procesos muy complicados que exigen el comportamiento diario. Cuando se analiza series de corte transversal el supuesto típico es que una variable económica y_t es generada por una distribución de probabilidad paramétrica del tipo $f(y/\theta)$, donde (θ) es un vector real de dimensión k "idéntico para todos los individuos en todo instante de tiempo". Este supuesto puede no ser realista en el caso de datos de panel; es más ignorar la heterogeneidad en los intercepto y/o en las pendientes es una que puede ser errada.

- **Sesgo de selección:**

Otra fuente de sesgo que se encuentra con frecuencia en datos de corte transversal y de paneles de datos es que la muestra puede no haber sido extraída de manera aleatoria de una población lo cual es poco frecuente en series de tiempo. Como consecuencia de ello se puede tener (de Arce & Mahía, 2007):

- Amplificación del efecto de errores de medida asociados a datos de encuestas.
- Falta de representatividad de la muestra debido a:
  - ✓ Desgaste muestral
  - ✓ No aleatoriedad de las observaciones

Ejemplos de este tipo de limitaciones se encuentran en: La cobertura de la población de interés, porcentajes de respuesta, preguntas confusas, distorsión deliberada de las respuestas, etc.

III. DISTINTAS FORMULACIONES DE UN MODELO CON DATOS DE PANEL

La fuente de variación muestral afecta la formulación y estimación de muchos modelos econométricos. Los tipos de modelos que se consideran en el caso de paneles de datos son (de Arce & Mahía, 2007):

i. Modelos con pendientes constantes en el tiempo y a nivel de cada individuo, pero los interceptos varían entre los individuos (Greene, 1999).

$$y_{it} = \alpha_i + \sum_{k=1}^{K} \beta_k x_{k,it} + \varepsilon_{it} \qquad \textbf{7.}$$

O en forma compacta

$$y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it} \qquad \textbf{8.}$$

ii. Modelos con pendientes constantes pero los intercepto varían entre los individuos y en el tiempo

$$y_{it} = \alpha_{it} + \sum_{k=1}^{K} \beta_k x_{k,it} + \varepsilon_{it} \qquad \textbf{9.}$$

O en forma compacta

$$y_{it} = \alpha_{it} + \beta' x_{it} + \varepsilon_{it} \qquad \textbf{10.}$$

iii. Todos los parámetros varían entre individuos pero se mantienen constantes en el tiempo.

$$y_{it} = \alpha_i + \sum_{k=1}^{K} \beta_{ki} x_{k,it} + \varepsilon_{it} \qquad \textbf{11.}$$

O en forma compacta:

$$y_{it} = \alpha_i + \beta_i' x_{it} + \varepsilon_{it} \qquad \textbf{12.}$$

iv. Todos los parámetros varían entre los individuos y en el tiempo

$$y_{it} = \alpha_{it} + \sum_{k=1}^{K} \beta_{k,it} x_{k,it} + \varepsilon_{it} \qquad \textbf{13.}$$

O en forma compacta

$$y_{it} = \alpha_{it} + \beta'_{it} x_{it} + \varepsilon_{it} \qquad \textbf{14.}$$

Entonces dentro de esta metodología en total tendríamos cuatro modelos.

Dónde:

$\beta_{k,it} = \beta_{i,t}^{k}$ Que son las distintas formas estándar que usan los autores para denotar a las pendientes lineales o coeficientes. En base a ello podemos formular la siguiente matriz de coeficientes para el modelo más general.

$\beta_{k,it} = \beta_{i,t}^{k}$: La lectura es la siguiente. Es el coeficiente asociado a la observación i-ésima de la variable k-ésima en el periodo temporal t.

Entonces para la observación i tenemos la siguiente matriz de coeficientes asociados a las variables explicativas.

$$\beta_i' = \begin{bmatrix} \beta_{i1}^{1} & \beta_{i1}^{2} & . & . & . & \beta_{i1}^{K-1} & \beta_{i1}^{K} \\ \beta_{i2}^{1} & \beta_{i2}^{2} & . & . & . & \beta_{i2}^{K-1} & \beta_{i2}^{K} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ \beta_{iT-1}^{1} & \beta_{iT-1}^{2} & . & . & . & \beta_{iT-1}^{K-1} & \beta_{iT-1}^{K} \\ \beta_{iT}^{1} & \beta_{iT}^{2} & . & . & . & \beta_{iT}^{K-1} & \beta_{T}^{K} \end{bmatrix} \qquad \textbf{15.}$$

$$\beta_i' = \begin{bmatrix} \beta_{i1} \\ \beta_{i1} \\ . \\ . \\ . \\ \beta_{iT-1} \\ \beta_{iT-2} \end{bmatrix} \qquad \textbf{16.}$$

Al mismo tiempo, los modelos del (i) al (iv) son clasificados dependiendo de si se asume que los coeficientes son aleatorios (random) o fijos (fixed), pero los modelos con pendientes constantes e interceptos variables (esto es modelos (i) y (ii) son los más usados al estimar modelos con datos en panel, puesto que proporcionan alternativas simples pero razonablemente amplias al supuesto de que los parámetros que son comunes a todos los agentes económicos.

Análisis de Covarianza: *(GREENE, 1999)*

- En el análisis econométrico, se asume que Y corresponde a realizaciones aleatorias de algún experimento con una cierta distribución de probabilidad asociada, condicionada a un vector de características y parámetros, $f(y/x,\theta)$

- Un modelo de datos de panel involucra a un total de n individuos o unidades económicas, con información a lo largo de T periodos, para la variable endógena y_{it} y para las variables exógenas $X_{k,it}$, donde i es el número de individuos, t es el instante de tiempo, k es el número de características o variables que lo distinguen entre sí.

- Entonces uno de los objetivos primordiales de los modelos de datos de panel es usar toda la información disponible para realizar las inferencias sobre θ. Generalmente se postula que Y es una función lineal de X; sin embargo, para correr una regresión mediante OLS con las NT observaciones es necesario asumir que los parámetros de la regresión toman valores comunes a través de todas de todas las unidades económicas o individuos para todo periodo de tiempo. Si este supuesto no es válido, la estimación OLS puede conducir a inferencias erróneas.

- Lo anterior significa que el primer paso hacia una explotación completa de los datos, es testear si los parámetros que caracterizan a la variable aleatoria Y, permanecen constantes entre los individuos y en el tiempo. El procedimiento recomendado es un "análisis de varianza".

- El análisis de varianza está enfocado en una categoría particular de hipótesis lineales que estipulan que el valor esperado de una variable aleatoria depende sólo de la clase a la cual los individuos considerados pertenecen; sin embargo el procedimiento excluye los test relacionados con regresiones. El análisis de covarianza, por su parte, tiene un carácter mixto involucrando a las variables genuinamente exógenas, como en el caso de los modelos de regresión, y al mismo tiempo permite que las verdaderas relaciones para cada individuo dependan de la clase a la cual este pertenece, tal como lo hacen los modelos de análisis de varianza.

- Con el objeto de considerar los efectos de factores tanto cualitativos como cuantitativos, se propone un modelo lineal como:

$$y_{it} = \alpha_{it} + \beta' x_{it} + \varepsilon_{it} \qquad \textbf{17.}$$

Y el procedimiento de testeo es (de Arce & Mahía, 2007):

(1) Probar si los $\alpha' s$ y los $\beta' s$ son o no simultáneamente homogéneos entre los distintos individuos y entre los distintos periodos de tiempo.

(2) Probar si los $\beta' s$ son o no los mismos colectivamente.

(3) Testear si los $\alpha' s$ son o no los mismos colectivamente.

En el análisis de covarianza de datos de panel se asume que los parámetros son constantes en el tiempo, pero varían entre individuos, lo que circunscribe el análisis a modelos del tipo (iii). Por lo tanto, es factible postular un modelo de regresión diferente para cada individuo. Esto es, el modelo a considerar es:

MODELO NO RESTRINGIDO (UNRESTRICTED) (Beltrán Barco, 2003)

A continuación analizamos las versiones de panel frecuentemente utilizadas:

$$y_{it} = \alpha_i + \beta'_i x_{it} + \varepsilon_{it} \qquad \textbf{18.}$$

Conocido como el modelo no restringido, Donde se plantean tres tipos de restricciones:

(i) Los coeficientes de las variables explicativas o pendientes son idénticos, pero no los interceptos (individual mean):

$$y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it} \quad \mathbf{19.}$$

Las diversas hipótesis alternas correspondientes, Hi:

(ii) Los interceptos son los mismos pero no las pendientes

H2: $$y_{it} = \alpha + \beta'_i x_{it} + \varepsilon_{it} \quad \mathbf{20.}$$

(iii) Los interceptos y las pendientes son los mismos entre individuos (pooled regresión):

H3: $$y_{it} = \alpha + \beta' x_{it} + \varepsilon_{it} \quad \mathbf{21.}$$

Considerando el modelo no restringido, el estimador OLS de β_i y de α_i están dadas por (Greene, 1999):

$$\hat{\beta}_i = W_{xx,i}^{-1} W_{xy,i} \quad \mathbf{22.}$$

$$\hat{\alpha}_i = \bar{y}_i - \beta' x_{it}$$

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it} \; ; \; \bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it} \qquad 23$$

$$W_{xx,i} = \sum_{t=1}^{T}\left(x_{it} - \bar{x}_i\right)\left(x_{it} - \bar{x}_i\right)' \; ;$$

$$W_{xy,i} = \sum_{t=1}^{T}\left(x_{it} - \bar{x}_i\right)\left(y_{it} - \bar{y}_i\right)' \qquad 23.$$

$$W_{yy,i} = \sum_{t=1}^{T}\left(y_{it} - \bar{y}_i\right)^2 \qquad 24.$$

Donde la suma de residuos al cuadrado del i-ésimo grupo es $SRC_i = W_{yy,i} - W'_{xy,i} W^{-1}_{xx,i} W_{xy.i}$, por lo que la suma no restringida de los residuos al cuadrado es $S_1 = \sum_{i=1}^{n} SCR_i$.

Si tomamos el modelo H1 (Denominado modelo con corrección de media individual), los estimadores serán:

$$\hat{\beta}_W = W_{xx}^{-1} W_{xy} \; ; \; \hat{\alpha}_i = \bar{y}_i - \hat{\beta}_w \bar{x}_i \qquad 25.$$

$$W_{xx} = \sum_{i=1}^{N} W_{xx,i} \; ; \; W_{xy} = \sum_{i=1}^{n} W_{xy,i} \quad W_{yy} = \sum_{i=1}^{N} W_{yy,i} \qquad 26.$$

En este caso la sumatoria de los residuos al cuadrado es

$$S_2 = W_{yy} - W_{xy}^{'} W_{xx} W_{xy} \quad \text{27.}$$

Finalmente, considerando ahora el modelo H3 (pooled model), los estimadores son:

$$\hat{\beta}_W = T_{xx}^{-1} T_{xy} \qquad \hat{\alpha} = \bar{y}_i - \hat{\beta}^{'} \bar{x}_i \quad \text{28.}$$

$$T_{xx} = \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{it} - \bar{x} \right) \left(x_{it} - \bar{x} \right)^{'} \quad \text{29.}$$

$$T_{xy} = \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{it} - \bar{x} \right) \left(y_{it} - \bar{y} \right)^{'} \quad \text{29.}$$

$$\bar{x} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it} \; ; \qquad \bar{y} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} y_{it} \quad \text{30.}$$

Y la suma total de residuos al cuadrado es:

$$S_3 = T_{yy} - T'_{xy} T_{xx}^{-1} T_{xy} \quad \text{31.}$$

Asumiendo que $\varepsilon_{it} \overset{iid}{\sim} N(0, \sigma_{\varepsilon}^2)$ a lo largo de i y de t, entonces es posible usar un test-F para testear las restricciones impuestas por “el modelo con corrección de medias individuales” y para el “pooled model”. Por ejemplo,

(a) La hipótesis de interceptos heterogéneos pero pendientes homogéneas puede formularse como el modelo no restringido sujeto a (N-1)k restricciones lineales:

$$y_{it} = \alpha_i + \beta'_i x_{it} + \varepsilon_{it} \quad \textbf{32.}$$

$$H^1{}_0 : \beta_1 = \beta_2 = \ldots = \beta_N \quad \textbf{33.}$$

(b) La hipótesis de interceptos y pendientes comunes pueden visualizarse como el modelo restringido sujeto a (N-1)(K+1) restricciones lineales:

$$y_{it} = \alpha_i + \beta'_i x_{it} + \varepsilon_{it} \quad \textbf{34.}$$

$$H^3{}_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_N \quad \textbf{35.}$$

$$\beta_1 = \beta_2 = \ldots = \beta_N \quad \textbf{36.}$$

Luego la aplicación de un test de análisis de covarianza es equivalente a un test de hipótesis basado en la suma de residuos al cuadrado provenientes de los resultados de un modelo de regresión lineal. Así, por ejemplo,para testear $H^3{}_0$ se utiliza:

$$F_3 = \frac{(S_3 - S_1) \big/ ([N-1][k+1])}{S_1 \big/ [NT - N(k+1)]} \approx F\left[(N-1)(k+1), nT - N(k+1)\right]$$

Por lo que si F3 no es significativa, entonces lo que conviene es agrupar los datos y estimar una ecuación de regresión agrupada (pooled regresión model). Por su parte, si F3 es significativamente grande, se debe realizar un esfuerzo adicional para detectar si la no homogeneidad puede atribuirse a las pendientes heterogéneas o interceptos heterogéneos.

Bajo la hipótesis de interceptos heterogéneos pero pendientes homogéneas (H1), el estadígrafo F es:

$$H_1 : \beta_1 = \beta_2 = \ldots = \beta_N \qquad \textbf{37.}$$

$$F_1 = \frac{(S_2 - S_1) \big/ ([n-1][k])}{S_1 \big/ [nT - n(k+1)]} \approx F\left[(n-1)k, nT - n(k+1)\right] \qquad \textbf{38.}$$

Y si F1 es significativa, se detiene la secuencia del testeo. Si F1 es no significativa, entonces podemos determinar la magnitud en la cual puede surgir la no

Homogeneidad en los interceptos.

Si H1 no es rechazada, también se puede aplicar un test condicional para la homogeneidad en los intercepto:

$$H_4 : \alpha_1 = \alpha_2 = \ldots . \alpha_N \qquad \textbf{39.}$$

Dado que $\beta_1 = \beta_2 = \ldots . \beta_N$ en cuyo caso la suma de cuadrados residuales no restringida es S2, mientras que la restringida es S3, y el estadígrafo F es:

$$F_4 = \frac{(S_3 - S_2) / N-1}{S_2 / \left[N(T-1)-k\right]} \approx F\left[N-1, N(T-1)-k\right] \qquad 40.$$

Alternativamente, podemos asumir que los coeficientes son constantes a través de los individuos en un instante dado, pero varían entre periodos; esto implica que podemos postular regresiones separadas para cada sección de corte transversal:

$$y_{it} = \alpha_t + \beta_t' x_{it} + \varepsilon_{it} \qquad 41.$$

Transformándose este último en nuestro modelo no restringido. Es posible realizar un análisis de covarianza análogo al anterior para este modelo, con el fin de testear la homogeneidad en los parámetros de corte transversal a lo largo de los distintos periodos de tiempo. Por ejemplo, podemos testear la homogeneidad global, planteando la siguiente hipótesis nula:

$$H_0^{3'} : \alpha_1 = \alpha_2 = \ldots \alpha_T \qquad 42.$$

$$\beta_1 = \beta_2 = \ldots . \beta_T \qquad 43.$$

Y el estadígrafo apropiado es:

$$F_3' = \frac{(S_3 - S_1') / \left[(T-1)(k+1)\right]}{S_1' \left[nT - T(k+1)\right]} \approx F\left[(T-1)(k+1), nT - T(k+1)\right] \qquad 44.$$

Y para testear la hipótesis de interceptos heterogéneos, pero pendientes homogéneas, dada por:

$$H_1^{'} : \alpha_1 \neq \alpha_2 \neq \ldots\alpha_T \qquad \textbf{45.}$$

$$\beta_1 = \beta_2 = \ldots.\beta_T \qquad \textbf{46.}$$

El estadígrafo es:

$$F_1^{'} = \frac{\left(S_2^{'} - S_1^{'}\right) / \left[(T-1)(k)\right]}{S_1^{'}\left[nT - T(k+1)\right]} \approx F\left[(T-1)K, nT - T(k+1)\right] \qquad \textbf{47.}$$

Para testear la hipótesis de interceptos homogéneos condicionados a la existencia de pendientes homogéneas:

$$H_0^{4'} : \alpha_1 = \alpha_2 = \ldots\alpha_T \qquad \textbf{48.}$$

Cuyo estadígrafo es:

$$F_1^{'} = \frac{\left(S_3^{'} - S_2^{'}\right) / \left[(T-1)\right]}{S_2^{'}\left[T(n-1) - k\right]} \approx F\left[(T-1), T(n-1) - k\right] \qquad \textbf{49.}$$

IV. ¿CUÁNDO UTILIZAR DATOS DE PANEL? (Labra, 2014)

Muchos trabajos de investigación en la última década han venido aplicando la metodología de datos de panel. Esto se debe en parte, al gran avance que ha existido en las bases de datos, las cuales se han elaborado recogiendo cada vez más datos de individuos a lo largo del tiempo.

Previos trabajos que han utilizado regresiones lineales, habían sido abordados usando técnicas de series de tiempo y de sección cruzada. Una extensión a las técnicas anteriores es el Pool de datos, según la cual cada individuo, en un momento de tiempo, es una observación.

El desarrollo de técnicas de datos de panel, por el contrario, puede tratar en forma independiente el conjunto de datos de un individuo en el tiempo, lo que se conoce como efectos individuales (α_i)

Como se observa en la figura 1, un conjunto de datos pueden ser analizados de diferentes formas. La figura 1A muestra una serie de observaciones analizadas, mientras que la figura 1B, presenta un análisis de regresión lineal, sin diferenciar individuos, es decir como un Pool de datos.

Por su parte, la figura 1C muestra como el conjunto de datos está compuesto en la realidad por tres subgrupos (individuos observados a través del tiempo) y que por tanto, podría llevarse a cabo otro tipo de análisis que considere esta condición.

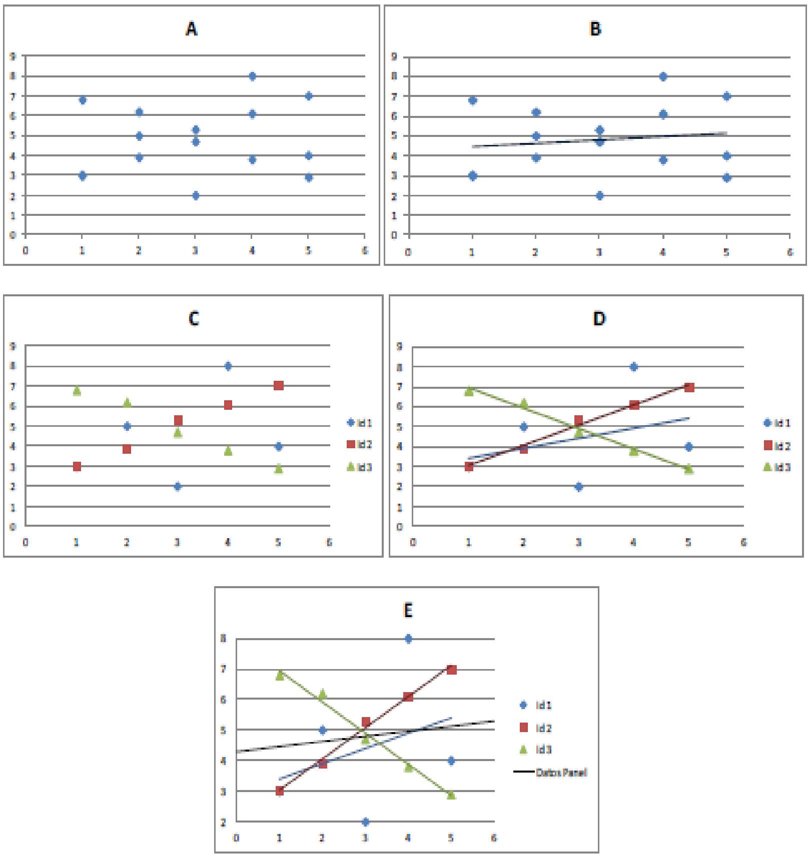
La figura 2D muestra las regresiones para cada subgrupo o individuo (individuo 1, 2 y 3), de lo que se desprende que cada uno posee un comportamiento diferente y que debe ser desarrollado teniendo en cuenta esta particularidad.

Finalmente, en la figura 2E se presenta una regresión lineal teniendo en cuenta los efectos individuales. La función final para el conjunto de individuos (línea continua de color negro) es totalmente distinta a la que obtendríamos si el análisis no se hubiese hecho a través de las técnicas de panel (figura 2B).

Como conclusión de este ejemplo gráfico, se puede deducir que es importante tener en cuenta los efectos individuales cuando estos existen, ya que el análisis y sus resultados pueden variar al usar una u otra técnica.

Cabe mencionar que dentro de esta última técnica, datos de panel, existen dos grandes métodos: Paneles Estáticos y Paneles Dinámicos, cuya principal diferencia radica en la capacidad y forma de tratar la endogeneidad de las variables, como se verá en los capítulos siguientes.

Figura 2. Esquema Gráfico de los efectos individuales (αi)



V. REGRESIÓN SIMPLE CON INTERCEPTOS VARIABLES Y EL MODELO DE COMPONENTES DE ERROR

Cuando la hipótesis de homogeneidad global es rechazada para el panel de datos, una forma simple de tomar en cuenta la heterogeneidad a través de individuos y/o del tiempo es el uso de un modelo de interceptos variables del tipo (Beltrán Barco, 2003):

$$y_{it} = \alpha_i + \sum_{k=1}^{n} \beta_k x_{k,it} + \varepsilon_{it} \tag{50.}$$

O del tipo

$$y_{it} = \alpha_{it} + \sum_{k=1}^{n} \beta_k x_{k,it} + \varepsilon_{it} \tag{51.}$$

El supuesto básico de estos modelos es que, condicionada a variables explicativas observadas, los efectos de todas las variables omitidas(o excluidas) son originados por tres tipos de variables:

- Variables individuales, invariantes en el tiempo (α_i): sexo, administración, capacidad empresarial, etc.
- Variables periódicas, invariantes entre individuos (λ_t): precios, tasas de interés, etc.
- Variables individuales, variables en el tiempo (Zit): políticas de las empresas, ventas, stock de capital, etc.

Los modelos de interceptos variables asumen que los efectos de una variedad de variables no constantes en el tiempo omitidas, son cada una de ellas individualmente no importantes pero colectivamente relevantes, y poseen la propiedad de una variable aleatoria que no está correlacionada (o es independiente) a todas las demás variables incluidas y excluidas del modelo.

Por otro lado, debido a que las restantes variables omitidas permanecen constantes en el tiempo para una unidad de corte transversal dada, o son las mismas para todas las unidades de corte transversal en un instante dado de tiempo, o son una combinación de ambas, por lo que pueden ser incorporadas dentro del intercepto de un modelo de regresión para así permitir explícitamente la posibilidad de heterogeneidad individual y/o en el tiempo contenida en los datos temporales de corte transversal. Esto es, el modelo sería (Mayorga M. & Muñoz S., 2000):

$$y_{it} = \mu + \sum_{k=1}^{n} \beta_{ki} x_{k,it} + v_{it} \qquad \textbf{52.}$$

Y si $v_{it} = \alpha_i + \lambda_t + \varepsilon_{it}$ **53.**

Dónde:

α_i : Componente aleatorio específico al individuo i.

λ_t : Componente aleatorio especifico al tiempo t

ε_{it} : Componente aleatorio

Entonces:

$$y_i = \mu + \alpha_i + \lambda_t + \beta_i{}' x_i + \varepsilon_{it} \qquad \textbf{54.}$$

Y en consecuencia:

$$y_{it} = \mu + \alpha_i + \lambda_t + \sum_{k=1}^{n} \beta_{ki} x_{k,it} + \varepsilon_{it} \qquad \textbf{55.}$$

Si corresponde al modelo general, asumiendo que los $x_{k,it}$ son no estocásticos.

$\sigma^2{}_\alpha ; \sigma^2{}_\lambda ; \sigma^2_\varepsilon$ Son independientes (incorrelación no implica independencia).

$$v_{it} = \alpha_i + \lambda_t + \varepsilon_{it} \qquad \textbf{56.}$$

$$E(v_{it}) = E(\alpha_i) = E(\lambda_t) = E(\varepsilon_{it}) = 0 \qquad \textbf{57.}$$

$$V(v_{it}) = \sigma^2{}_\alpha + \sigma^2{}_\lambda + \sigma^2_\varepsilon \qquad \textbf{58.}$$

Dos alternativas:

i) $Cov(x_{it}, \varepsilon_{it}) \neq 0 \Rightarrow Cov(x_{it}, v_{it}) \neq 0 \Rightarrow$ Inconsistencia.

ii) $Cov(x_{it}, v_{it}) = 0 \quad E(X' \alpha_i) = 0$

La consistencia se refiere que a medida que aumenta el tamaño de la muestra $N \to \infty$ los estimadores convergen a sus verdaderos valores poblacionales.

La mayoría de las aplicaciones con datos de panel utilizan el modelo de componente de error conocido como "one way" para el cual $\lambda_t = 0$ [5]. Las diferentes variantes para el modelo "one way" de componentes de errores surgen de los distintos supuestos que se hacen acerca del término α_i . Pueden presentarse tres posibilidades (Mayorga M. & Muñoz S., 2000)s:

El caso más sencillo es el que considera al $\alpha_i = 0$ o sea, no existe heterogeneidad no observable entre los individuos o firmas. Dado lo anterior, los ε_{it} satisfacen todos los supuestos del modelo lineal general, por lo cual el método de estimación de mínimos cuadrados clásicos produce los mejores estimadores lineales e insesgados.

La segunda posibilidad consiste en suponer a α_i un efecto fijo y distinto para cada firma. En este caso, la heterogeneidad no observable se incorpora a la constante del modelo.

La tercera alternativa es tratar a α_i como una variable aleatoria no observable que varía entre individuos pero no en el tiempo1

[1] Este tipo de análisis supone que no existen efectos no cuantificables que varíen en el tiempo pero no entre las unidades individuales de estudio. Existe además el modelo "two-way" en el cual el componente de error $\lambda_t \neq 0$ a través del cual se pretende capturar efectos temporales específicos (choques) que no están incluidos en la regresión. Para un mayor detalle puede consultarse a Balgati (1999), capítulo 3.

VI. MODELO BÁSICO DE HETEROGENEIDAD:

$$y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it} \quad 59.$$

Donde α_i es la diferenciación atemporal de los individuos. Si todos los α_i son iguales a una constante s, por ejemplo, se estaría en el caso de POOL DATA esto ya se mencionó anteriormente. Cuando se da esta situación entonces ignoramos la estructura de un panel data en los datos. Apilamos la data y obtenemos:

$$Y = X \beta + \varepsilon \quad 60.$$

$$\varepsilon_{it} \approx iid(0, \sigma^2) \forall i \wedge t \quad 61.$$

Esto es, las observaciones están serialmente incorrelacionadas; a través de los individuos y a través del tiempo, además de que los errores son homocedásticos.

La estimación de este modelo es directa. Los estimadores son eficientes debido a que la data fue apilada de forma tal que se puede aplicar MCO. Pero esta no es la estimación que se usa por lo comentado anteriormente.

Recordando dos procedimientos adicionales para estimar el modelo en un sistema de datos de panel: uno de ellos implica el reconocimiento de que las variables omitidas pueden generar cambios en los interceptos ya sea a través del tiempo o entre unidades de corte transversal, en este caso el modelo de efectos fijos que trata de aproximar estos cambios con variables dummy; el otro modelo es el de efectos aleatorios, que trata de capturar estas diferencias a través del componente aleatorio del modelo.

VII. MODELO DE EFECTOS FIJOS

Como se indicó brevemente, una posibilidad es explicar los datos con el modelo de efectos fijos considera que existe un término constante diferente para cada individuo, y supone que los efectos individuales son independientes entre sí.

Con este modelo se considera que las variables explicativas afectan por igual a las unidades de corte transversal y que éstas se diferencian por características propias de cada una de ellas, medidas por medio del intercepto. Es por ello que los N interceptos se asocian con variables dummy con coeficientes específicos para cada unidad, los cuales se deben estimar (Beltrán Barco, 2003).

Entonces, este modelo, la diferencia entre las observaciones se captura por la constante, para cada momento t.

En el modelo de efectos fijos, los efectos individuales i (i = 1,....., N) se tratan como coeficientes desconocidos que se estiman conjuntamente con[2].

$$y_i = i_T \alpha_i + x_i \beta + \varepsilon_i \qquad 62.$$

[2] Hay que tomar en cuenta que *i* no es un coeficiente que multiplica a α_i .

Dónde:

$$i_T = \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \\ 1 \\ 1 \end{bmatrix}_{TX1} \qquad 62.$$

Debe hacerse notar que en este modelo se presenta una pérdida importante de grados de libertad.

Para todo T:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{NTx1} = \begin{bmatrix} i & 0 & \dots & 0 \\ 0 & i & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & i \end{bmatrix}_{NTxN} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix}_{Nx1} + \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}_{NTxK} \beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}_{NTx1} \qquad \textbf{62.}$$

Lo que en términos vectoriales sería:

$$\underline{y} = \begin{bmatrix} d_1 & d_2 & \dots & d_N & X_k \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \varepsilon \qquad \textbf{63.}$$

Dónde:

$$d_i = \begin{cases} 1 \quad para \;\; las \;\; T \;\; observaciones \;\; del \;\; individuo \;\; i \\ 0 \quad para \;\; cualquier \;\; t \;\; de \;\; cualquier \;\; otra \;\; observación \end{cases}$$

En resumen:

$$y = D\alpha + \beta x + \varepsilon \qquad \textbf{64.}$$

Este modelo se denomina habitualmente como el modelo de mínimos cuadrados de variables ficticias (MCVF) (aunque la parte del nombre de "mínimos cuadrados" se refiere a la técnica que se utiliza habitualmente para estimarlo, no al modelo como tal).

Si N es pequeño, se estima con MCO, con K regresores en X y N columnas en D, como una regresión múltiple con (N+K) parámetros. Si N es grande (como en el caso de una encuesta), entonces se procede a utilizar regresiones particionadas, en las que el estimador es:

$\hat{\beta}_w = [x'M_d x]^{-1}[x'M_d y]$ Es un estimador que se obtiene de regresionar por MCO toda la data incluyendo las variables Dummys.

A este estimador se le conoce como estimador Intra(Whitin) de efectos fijos . Este tipo de estimadores estudian el modelo en desviaciones.

Donde $M_d = I_{NT} - D(D'D)^{-1}D$ La cual es una matriz simétrica o idempotente ó, matricialmente,

$$M_d = \begin{bmatrix} M_0 & & & \phi \\ & M_0 & & \\ & & \ddots & \\ \phi & & & M_0 \end{bmatrix} \quad 65.$$

y donde $M_0 = I_T - \frac{ii'}{T}$. Esto equivale a regresionar $M_d X$ y $M_d Y$, es decir,

$$(y_{it} - \bar{y}_i) \text{ vs. } (x_{it} - \bar{x}_i)$$

Es decir:

$$(y_{it} - \bar{y}_i) = (x_{it} - \bar{x}_i)\beta + \varepsilon_i \quad 66.$$

Dónde:

$$\bar{y}_i = \sum_{t=1}^{T} \frac{y_{it}}{T} \quad \text{y} \quad \bar{x}_i = \sum_{t=1}^{T} \frac{x_{it}}{T} \quad 67.$$

Este es el llamado Modelo LSDV (Least Squares Dummys Variables)

Ahora, para encontrar los parámetros de interés (α_i's):

$$\hat{\alpha}_i = \bar{y}_i - \hat{\beta}\,\bar{x}_i \quad 68.$$

La variancia del estimador:

$$Var(\hat{\beta}) = s^2\left[x' M_d x\right]^{-1} \quad donde: s^2 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{n}(y_{it} - \hat{\alpha} - \hat{\beta}x_{it})^2}{NT - N - k} \qquad \textbf{69.}$$

Y donde $NT - N - k$ es el número total de observaciones.

Para poder decidir si utilizar el Modelo LSDV o el Modelo de POOL DATA, se procede a realizar una prueba F ,ya vista anteriormente, para observar las diferencias entre los grupos, podemos contrastar la hipótesis de que los términos constantes son todos iguales, mediante un contraste F con la siguiente especificación:

H_o: El mejor modelo es un POOL de datos con constante única $(\alpha_i = \alpha)$

$$F_{(N-1, NT-N-k)} = \frac{(R_u^2 - R_{POOL}^2)/(N-1)}{(1-R_u^2)/(NT-N-k)} \qquad \textbf{70.}$$

Donde u indica el modelo no restringido y POOL indica el modelo agrupado, o restringido, con un único término constante para todos.

Bajo la hipótesis nula, el estimador eficiente coincide con mínimos cuadrados agrupados.

VIII. VARIANTE DEL MODELO LSDV: EL MODELO DE EFECTOS FIJOS TEMPORALES

El enfoque de mínimos cuadrados con variables artificiales se puede entender, también, para que incluya un efecto temporal específico.

El modelo de efectos fijos es un enfoque razonable cuando podemos estar seguros que las diferencias entre unidades se pueden interpretar como un desplazamiento paramétrico de la función de regresión. Este modelo podría interpretarse:

$$y_{it} = \alpha_i + \gamma_t + \beta' x_{it} + \varepsilon_{it} \qquad \textbf{71.}$$

Donde los $\beta's$ son para todos los individuos y controlados por los efectos individuales. Se deberá incluir T-1 variables dicotómicas. Uno de los efectos temporales debe dejarse fuera para evitar colinealidad perfecta. Si el número de variables es de demasiado grande para manejarlo mediante la regresión ordinaria, puede estimarse, también, utilizando la regresión particionada. Sin embargo hay una asimetría en esta formulación, ya que cada uno delos efectos de grupo es una constante específica de grupo, mientras que los efectos temporales son Referencias, es decir, comparaciones con un año base (el excluido). Una forma simétrica del modelo es

$$y_{it} = \mu + \alpha_i + \gamma_t + \beta' x_{it} + \varepsilon_{it} \qquad \textbf{72.}$$

Donde los efectos N y T, están incluidos pero las restricciones $\sum_i \alpha_i = \sum_t \gamma = 0$

se incorporan.

La ecuación particionada es de la siguiente forma:

$$\left[y_{it} - \bar{y}_i - \bar{y}_t + \bar{\bar{y}}\right] vs \left[x_{it} - \bar{x}_i - \bar{x}_t + \bar{\bar{x}}\right] \qquad \textbf{73.}$$

Dónde:

$$\bar{y}_t = \sum_{i=1}^{N} \frac{y_{it}}{N} \qquad \textbf{74.}$$

$$\bar{\bar{y}} = \sum_{i=1}^{N} \sum_{t=1}^{N} \frac{y_{it}}{Nt} \qquad \textbf{75.}$$

Los parámetros de interés, en esta especificación son:

$$\begin{aligned} \hat{\alpha} &= (\bar{y}_i - \bar{\bar{y}}) - \hat{\beta}(\bar{x}_i - \bar{\bar{x}}) \\ \hat{\gamma} &= (\bar{y}_t - \bar{\bar{y}}) - \hat{\beta}(\bar{x}_t - \bar{\bar{x}}) \end{aligned} \qquad \textbf{76.}$$

Que miden los efectos individuales y efectos temporales, respectivamente.

Limitaciones del Modelo LSDV:

- Pérdida de Grados de Libertad.
- No se puede hacer inferencia fuera de la muestra

Cabe destacar, sin embargo, que su principal ventaja es que no tiene que cuidarse que los efectos individuales estén no correlacionados con las X's, ya que por la especificación del modelo ello está asegurado.

IX. MODELO DE EFECTOS FIJOS

La generalización del modelo con interceptos y pendientes constantes para un panel de datos es introducir variables dummy para incorporar los efectos de aquellas variables omitidas que son específicas a las unidades individuales de corte transversal. Así por ejemplo, asumiendo que no existen efectos específicos al instante de tiempo $\lambda_t = 0$, el modelo es:

$$y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it} \qquad \textbf{77.}$$

Que corresponde al modelo de análisis de covarianza. En este modelo, ε_{it} representa los efectos de las variables explicativas omitidas que son particulares tanto a las unidades de corte transversal como a los periodos de tiempo considerados. Se asume que $\varepsilon_{it} \overset{iid}{\sim} N\left(0, \sigma_\varepsilon^2\right)$.

El estimador OLS de β es:

$$\hat{\beta}_{CV} = \left[\sum_{i=1}^{n}\sum_{t=1}^{T}\left(x_{it} - \bar{x}_i\right)\left(x_{it} - \bar{x}_i\right)'\right]^{-1}\left[\sum_{i=1}^{n}\sum_{t=1}^{T}\left(x_{it} - \bar{x}_i\right)\left(y_{it} - \bar{y}_i\right)'\right] \qquad \textbf{78.}$$

$$\hat{\alpha}_i = \bar{y}_i - x_i \hat{\beta}_{DV}$$

Que es el estimador OLS de β del modelo de covarianza.

Y dado que $\varepsilon_{it} \approx N(o, \sigma_\varepsilon^2)$, $\hat{\beta}_{CV}$ es MELI (mejor estimador lineal insesgado).

Por supuesto, para calcular no es necesario incluir variables dummy. Basta con calcular las medias x_i e y_i, de forma separada para cada unidad de corte tranversal. A continuación calcular $(y_i - \overline{y_i})e(x_i - \overline{x_i})$, y luego aplicar OLS al modelo transformado, cuya especificación es:

$$\left(y_{it} - \bar{y}_i\right) = \beta'\left(x_{it} - \bar{x}_i\right) + \left(\varepsilon_{it} - \bar{\varepsilon}_i\right) \qquad \textbf{79.}$$

Este procedimiento es equivalente a premultiplicar la ecuación

$y_{it} = i_T\alpha_i + \beta' x_{it} + \varepsilon_{it}$ Por $M_0 = I_t - \dfrac{ii'}{T}$, donde es un vector de unos, por lo que:

$$\hat{\beta}_{CV} = \hat{\beta}_W = \left[\sum_{i=1}^{n} x_i' M_0 x_i\right]^{-1}\left[\sum_{i=1}^{N} x_i' M_0 y_i\right] \qquad \textbf{80.}$$

Dado que el modelo $y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it}$ también se conoce como modelo de análisis de covarianza, el estimador de mínimos cuadrados con variables dummy se denomina estimador de covarianza. También recibe el nombre de estimador intragrupal (within estimador), puesto que solo se utiliza la variación al interior de cada grupo para calcular el estimador.

Como se ha señalado, el estimador $\hat{\beta}_{CV}$ es insesgado y también consistente cuando n y T tienden al infinito. Además la matriz de varianzas y covarianzas está dada por:

$$Var\left(\hat{\beta}_{CV}\right) = Var\left(\hat{\beta}_{W}\right) = \sigma_{\varepsilon}^{2}\left[\sum_{i=1}^{n} x_i ' M_0 x_i\right]^{-1} \qquad \textbf{81.}$$

por otro lado el estimador de α_i $\left(\hat{\alpha}_i = \bar{y}_i - \hat{\beta}_{DV}{}' \bar{x}_i\right)$ es insesgado, pero consistente sólo cuando T tiende al infinito.

X. MODELO DE EFECTOS ALEATORIOS

La decisión acerca de la estructura apropiada para el análisis, es decir, efectos fijos vs efectos aleatorios depende en parte de los siguientes aspectos:

a. Los objetivos del estudio

Si se desea hacer inferencias con respecto a la población, es decir que se trabaja con una muestra aleatoria, lo mejor es utilizar una especificación del tipo aleatoria. En caso de que el interés sea limitado a una muestra que se ha seleccionado a conveniencia o bien que se está trabajando con la población, la estimación de efectos fijos será la correcta.

Adicionalmente, si el interés del estudio particular está puesto en los coeficientes de las pendientes de los parámetros, y no tanto en las diferencias individuales, se debería elegir un método que relegue estas diferencias y tratar la heterogeneidad no observable como aleatoria.

El modelo de efectos fijos se ve como un caso en que el investigador hace inferencia condicionada a los efectos que ve en la muestra es decir marginaliza la función de probabilidad conjunta. El de efectos aleatorios se ve como uno en el cual el investigador hace inferencia condicional o marginal respecto a una población.

Se deja al investigador que decida si hace inferencia con respecto a las características de una población o solo respecto a los efectos que están en la muestra.

b. El contexto de los datos, es decir, cómo fueron obtenidos y el entorno de donde provienen.

Con el método de efectos fijos la heterogeneidad no observable se incorpora en la ordenada al origen del modelo y con la de efectos aleatorios, como ya se mencionó, se incorporan en el término de error, por lo cual lo que se modifica es la varianza del modelo.

Emplear un modelo de efectos fijos o aleatorios genera diferencias en las estimaciones de los parámetros en los casos en que se cuenta con t pequeño y N grande. En estos casos debe hacerse el uso más eficiente de la información para estimar esa parte de la relación de comportamiento contenida en las variables que difieren sustancialmente de un individuo a otro.

c. Número de datos disponibles.

El método de efectos fijos presenta el problema de que el uso de variables "Dummies" no identifica directamente qué causa que la regresión lineal cambie en el tiempo y en los individuos. Además, esto implica la pérdida de grados de libertad. Asimismo, deberán tomarse consideraciones con respecto a la estructura de los datos con que se cuente, dado que si la N es grande pero si se tiene un T pequeño, podría ser que el número de parámetros de efectos fijos sea muy grande en relación con el número de datos disponibles, con parámetros poco confiables y una estimación ineficiente. Algunas investigaciones han demostrado que el emplear modelos de efectos fijos produce resultados significativamente diferentes al de efectos aleatorios cuando se estima una ecuación usando una muestra de muchas unidades de corte transversal con pocos periodos de tiempo (629 individuos para 6 periodos, por ejemplo).

Sea: $$y_{it} = \beta \ x_{it} + \varepsilon_{it} \qquad \textbf{82.}$$

Donde el número de individuos es grande, y el número de periodos temporales es pequeño. Entonces haremos uso de la siguiente estructura del error del término de perturbación:

$$\varepsilon_{it} = \alpha_i + \eta_t + v_{it} \qquad \textbf{83.}$$

α_i : Llamado efecto individual o específico a cada grupo. Representa los efectos no observables que difieren entre las unidades de estudio pero no en el tiempo, que generalmente se los asocia a la capacidad empresarial, Por ejemplo: Conocimiento, habilidades.

v_{it} : Componente aleatorio.

η_t : Llamado efecto temporal. (Tipo de cambio, tasa de interés, inflación). Se le identifica con efectos no cuantificables que varían en el tiempo pero no entre las unidades de estudio.

La mayoría de las aplicaciones con datos de panel utilizan el modelo de componente de error conocido como "one way" para el cual $\eta_t = 0$. Las diferentes variantes para el modelo "one way" de componentes de errores surgen de los distintos supuestos que se hacen acerca del término α_i. Pueden presentarse tres posibilidades:

· El caso más sencillo es el que considera al $\alpha_i = 0$, o sea, no existe heterogeneidad no observable entre los individuos o firmas. Dado lo anterior, los ε_{it} satisfacen todos los supuestos del modelo lineal general, por lo cual el método de estimación de mínimos cuadrados clásicos produce los mejores estimadores lineales e insesgados.

· La segunda posibilidad consiste en suponer a α_i un efecto fijo (modelo que ya se explicó en secciones previas) y distinto para cada firma. En este caso, la heterogeneidad no observable se incorpora a la constante del modelo.

· La tercera alternativa es tratar a α_i como una variable aleatoria no observable que varía entre individuos pero no en el tiempo.

Si se da el caso en el que $Cov(x_{it};\alpha_i)\neq 0$ entonces en tal caso no hay manera de estimar los α_i de manera independiente ni siquiera con las DUMMY.

Por lo tanto la única manera de estimar los α_i es conjuntamente con el error v_{it} : (componente aleatorio).

Por este motivo el modelo de efectos aleatorios tiene un término de error con dos componentes:

$$\varepsilon_{it} = \alpha_i + v_{it} \tag{84.}$$

$$E(v_{it}) = 0 \tag{85.}$$

$$E(v^2{}_{it}) = \sigma_v^2 \tag{86.}$$

En esta formulación se ignora el efecto temporal la misma que puede o no estar correlacionada con las variables explicativas.

El modelo de efectos aleatorios es más apropiado cuando se trabaja con una muestra reducida de una gran población porque se puede inferir el comportamiento de individuos fuera de la muestra (en el caso anterior los efectos de cada individuo eran fijos, propios).

El modelo de efectos aleatorios tiene la siguiente especificación:

$$y_{it} = \beta \; x_{it} + \varepsilon_{it} \qquad \textbf{87.}$$

$$y_{it} = \alpha_i + \beta \; x_{it} + v_{it} \qquad \textbf{88.}$$

Donde $\alpha_i + v_{it}$ es el componente del error de la ecuación y que no está correlacionado con las X's.

Esta especificación nos permite capturar el hecho de que dos observaciones del mismo individuo son más parecidas que dos que pertenecen a diferentes individuos. Las propiedades, de dicho componente, son las siguientes:

Es importante resaltar este modelo se distingue del modelo de efectos fijos es que el efecto individual α_i está incorrelacionado con x_{it}. Esta es la condición de ortogonalidad con este supuesto los estimadores MCO son asintóticamente insesgados.

El problema es dual cuando el modelo verdadero es el de efectos aleatorios:

1. MCO producirá estimadores consistentes de β pero las desviaciones estándar subestimados.

2. MCO no es eficiente comparado con los procedimientos MCG.

Efectos aleatorios es una forma de tratar con el hecho de que T observaciones sobre n individuos no son lo mismo que nT diferentes individuos. La solución es directa. Primero, derivamos un estimador de la matriz de covarianzas del término de error. Segundo, usamos esta estructura en nuestro estimador de β.

Sea la naturaleza aleatoria del término del error:

$$E\left[v_{it}\right] = E\left[\alpha_i\right] = 0$$
$$E\left[v_{it}\right]^2 = \sigma_v^2 \quad E\left[\alpha_i\right]^2 = \sigma_\alpha^2$$
$$E\left[v_{it}'\alpha_j\right] = 0 \quad \forall\, i, t, j$$
$$E\left[v_{it}'v_{js}\right] = 0 \quad \forall\, t \neq s \quad e \quad i \neq j$$
$$E\left[\alpha_i'\alpha_j\right] = 0 \quad \forall\, i \neq j$$

89.

$$\varepsilon_{it} = \alpha_i + v_{it}$$

Para las T observaciones de cada individuo:

$$\varepsilon_{it} = v_{it} + \mu_i$$
$$\varepsilon_i = \left[\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{it}\right]'$$

Donde ε_{it}, este se denomina con frecuencia modelos de componentes de error.

Entonces:

$$E\left[\varepsilon_{it}\right]^2 = \sigma_v^2 + \sigma_\alpha^2$$
$$E\left[\varepsilon_{it}\varepsilon_{is}\right] = \sigma_\alpha^2$$
$$E\left[\varepsilon_{it}\varepsilon_{jt}\right] = 0 \qquad , t \neq s$$

$$i \neq j$$

Donde se aprecia que la correlación temporal depende sólo del efecto aleatorio individual.

Para las T observaciones del individuo i, $\Omega = E[\varepsilon_i \varepsilon_i']$ donde:

$$\Omega = \begin{bmatrix} \sigma_v^2 + \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_v^2 + \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_v^2 + \sigma_\alpha^2 \end{bmatrix} = \sigma_v^2 I + \sigma_\alpha^2 ii' \qquad \textbf{90.}$$

Donde i es un vector columna Tx1 de unos. Como las i y j observaciones son independientes, la matriz de varianzas y covarianzas de los errores para las nxT observaciones:

$$V = \begin{bmatrix} \Omega & & \phi \\ & \Omega & \\ & & \ddots & \\ \phi & & & \Omega \end{bmatrix} \qquad \textbf{91.}$$

En el contexto de datos de panel, existe una distinción tradicional entre "modelos de Efectos fijos" y "modelos de efectos aleatorios". En el modelo de efectos fijos, los efectos individuales α_i (i = 1,…..,N) se tratan como coeficientes desconocidos que se estiman conjuntamente con β.

En el modelo de efectos aleatorios, por el contrario, se supone que α_i es una variable aleatoria inobservable no correlacionada con las variables explicativas del modelo Xit, por lo que se incluye en el término de error:

$$\varepsilon_{it} = \alpha_i + v_{it} \tag{92.}$$

Por tanto, en el modelo de efectos aleatorios tenemos un término de error compuesto, por ε_{it} no correlacionado con las variables explicativas del modelo. Este modelo forma parte de los modelos de efectos compuestos.

En el trabajo empírico, muchas veces se estiman ambos modelos y se contrasta si los efectos son fijos o aleatorios. Sin embargo, esta estrategia no es apropiada, porque lo importante no es si los efectos individuales son fijos o aleatorios, porque de hecho en general tienen un carácter aleatorio.

El aspecto verdaderamente relevante, es conocer si dichos efectos individuales están o no correlacionados con las variables explicativas del modelo. (De hecho, el modelo de efectos fijos puede verse como un modelo de efectos aleatorios en que los efectos individuales sí están correlacionados con las variables explicativas del modelo).

La idea detrás de la estimación con parámetros incidentales (efectos fijos) responde a la necesidad de capturar las diferencias entre unidades de estudio que son capturadas por los represores empleados. El hecho de que se tenga un solo intercepto para cada individuo tiene como supuesto la heterogeneidad de la muestra.

Estimación

Mínimos Cuadrados Generalizados (cuando se conoce Ω)

Al estimar por MCG se obtienen estimaciones eficientes. Requerimos hallar el T que transforme la data. Así:

$$T = V^{-1/2} = I \otimes \Omega^{-1/2}$$

$$\Omega^{-1/2} = I - \frac{\theta i i'}{T} \quad \theta = 1 - \frac{\sigma_v^2}{(T\sigma_\alpha^2 + \sigma_v^2)^{1/2}} \qquad \textbf{93.}$$

Eso implica que la regresión particionada con la data transformada (premultiplicada por Ω -1/2) sería:

$$(y_{it} - \theta \bar{y}_i)\ vs.\ (x_{it} - \theta \bar{x}_i) \qquad \textbf{94.}$$

Si $\theta = 1$, entonces se está en el Modelo LSDV.

Una forma alternativa de estimación MCG consiste en formular la regresión de tres modos:

Modelo Original (a)[3]

$$y_{it} = \alpha + \beta \quad x_{it} + \varepsilon_{it} + \mu_i \qquad \textbf{95.}$$

[3] Modelo completo. Modelo dentro de las unidades (within the units). Modelo entre las unidades (between the units)

Modelo de Desviaciones de la media de cada i sobre el tiempo[8] (w):

$$(y_{it} - \bar{y}_i) = \beta(x_{it} - \bar{x}_i) + \varepsilon_{it} - \bar{\varepsilon}_i \tag{96.}$$

Modelo de cada individuo sobre T9 (b)

$$\bar{y}_i = \alpha + \beta\bar{x}_i + \bar{e}_i + \mu_i \tag{97.}$$

Además se definen las siguientes correlaciones:

Modelo t:

$$T_{xx}^{t} = \sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \overline{\overline{x}})(x_{it} - \overline{\overline{x}})$$

$$T_{xy}^{t} = \sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \overline{\overline{x}})(y_{it} - \overline{\overline{y}})$$ **97.**

Modelo w:

$$W_{xx} = \sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \overline{x}_i)(x_{it} - \overline{x}_i)$$

$$W_{xy} = \sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \overline{x}_i)(y_{it} - \overline{y}_i)$$ **98.**

Modelo b:

$$T_{xx}^{b} = \sum_{i=1}^{n}\sum_{t=1}^{T}(\overline{x}_i - \overline{\overline{x}})(\overline{x}_i - \overline{\overline{x}})$$

$$T_{xy}^{b} = \sum_{i=1}^{n}\sum_{t=1}^{T}(\overline{x}_i - \overline{\overline{x}})(\overline{y}_i - \overline{\overline{y}})$$ **99.**

Y se cumplen las siguientes propiedades:

$$T_{xx}^{t} = W_{xx} + T_{xx}^{b}$$
$$T_{xy}^{t} = W_{xy} + T_{xy}^{b}$$ **100.**

Para descomponer el estimador:

$$\hat{\beta}^{t} = \left[T_{xx}^{t}\right]^{-1} T_{xy}^{t}$$
$$\hat{\beta}^{t} = \left[W_{xx} + T_{xx}^{b}\right]^{-1}\left[W_{xy} + T_{xy}^{b}\right]$$ **101.**

Para los modelos w y b:

$\hat{\beta}^{w} = \left[W_{xx}\right]^{-1} T_{xy}^{w}$ → Que es el estimador del Modelo LSDV

$$\hat{\beta}^{b} = \left[T_{xx}^{b}\right]^{-1} T_{xy}^{b}$$

Lo anterior se puede rescribir como:

$$W_{xy} = W_{xx}\hat{\beta}^{w}$$
$$T_{xy}^{b} = T_{xx}^{b}\hat{\beta}^{b} \qquad \textbf{102.}$$

En consecuencia el estimador del modelo total sería:

$$\hat{\beta}^{t} = \left[W_{xx} + T_{xx}^{b}\right]^{-1}\left[W_{xx}^{w}\hat{\beta}^{w} + T_{xx}^{b}\hat{\beta}^{b}\right]$$
$$\hat{\beta}^{t} = F^{w}\hat{\beta}^{w} + F^{b}\hat{\beta}^{b} \qquad \textbf{103.}$$

Y donde $F^{w} = \left[W_{xx} + T_{xy}^{b}\right]^{-1} W_{xx}$. Las F^{w} y F^{b} son los pesos para explicar las variables.

Es posible demostrar que el estimador de MCG se puede escribir de la siguiente forma:

$$\hat{\beta}^{t}_{MCG}=F^{w}\hat{\beta}^{w}+F^{b}\hat{\beta}^{b}$$ y $$F^{w}=\left[W_{xx}+\lambda\ \ T^{b}_{xx}\right]^{-1}W_{xx}$$ donde

$$\lambda=\frac{\sigma^{2}_{v}}{\sigma^{2}_{v}+T\sigma^{2}_{\alpha}}=(1-\theta)^{2} \qquad \textbf{104.}$$

El proceso llevado acabo sirve para que el efecto de μ sea corregido en la matriz de variancias y covariancias.

Si $\lambda=1$ entonces, el estimador MCO es idéntico al MCG.

Si $\lambda=0$ entonces, se está frente al Modelo LSDV.

Mínimos Cuadrados Generalizados (cuando no se conoce Ω)

Estimación de $\hat{\sigma}^{2}_{\varepsilon_i}$

Para construir el estimador, es posible tomar información del Modelo w.

$$\hat{\sigma}^{2}_{\varepsilon_i}=\frac{\sum_{t=1}^{T}(\varepsilon_{it}-\bar{\varepsilon}_i)^2}{T-1} \qquad \textbf{105.}$$

$$\hat{s}^{2}_{\varepsilon_i}=\frac{\sum_{t=1}^{T}(\varepsilon_{it}-\bar{\varepsilon}_i)^2}{T-k-1}$$ Donde $T-K-1$ es el número de variables explicativas.

La Variancia total, en consecuencia, es:

$$\overline{s}_{\varepsilon}^{2} = \frac{s_{\varepsilon_i}^{2}}{N}$$

$$= \frac{1}{N}\sum_{i=1}^{n}\left[\frac{\sum_{t=1}^{T}(\varepsilon_{it}-\bar{\varepsilon}_i)^2}{T-k-1}\right]$$

$$= \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(\varepsilon_{it}-\bar{\varepsilon}_i)^2}{NT-Nk-N} \quad \textbf{106.}$$

Como se puede observar, la corrección por grados de libertad es excesiva (se asume que α y β son estimadores para cada unidad i). Pero los parámetros estimados son la n media de $\overline{Y}_i$ y las k pendientes ⇒ el estimador propuesto es:

$$\hat{\sigma}_{\varepsilon}^{2} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(\varepsilon_{it}-\bar{\varepsilon}_i)^2}{NT-N-N} \approx \sigma_e^2 \quad LSDV \quad \textbf{107.}$$

Estimación de $\hat{\sigma}_{\mu}^{2}$

Del modelo en medias:

$$e_{**_i} = \overline{y}_i - \alpha - \beta x_i$$
$$= \overline{e}_i + \mu_i$$

Analizando la varianza:

$$\sigma_{e_{**}}^{2} = \frac{e_{**}' e_{**}}{n-k} = \frac{\sigma_{e}^{2}}{T} + \sigma_{\mu}^{2} \qquad \textbf{108.}$$

En consecuencia:

$$\sigma_{\mu}^{2} = -\frac{\sigma_{e}^{2}}{T} + \sigma_{e_{**}}^{2} \qquad \textbf{109.}$$

MÍNIMOS CUADRADOS ORDINARIOS EN DESVIACIONES

Existe una forma más simple de estimar el modelo de efectos aleatorios, se denomina MCO en desviaciones. Los pasos a seguir son los siguientes:

Calcular los estimadores $\hat{\beta}^{w}$ y $\hat{\beta}^{b}$

Usar los residuos de cada estimación para hallar $\hat{\sigma}_{\alpha}^{2}$ y $\hat{\sigma}_{v}^{2}$

Estimar θ, donde $\hat{\theta} = \dfrac{\hat{\sigma}_{v}^{2}}{(T\hat{\sigma}_{\alpha}^{2} + \hat{\sigma}_{v}^{2})^{1/2}}$

Estimar MCO: $(\tilde{y}_{it} = y_{it} - \hat{\theta}\,\overline{y}_{i})$ vs. $(\tilde{x}_{it} = x_{it} - \hat{\theta}\,\overline{x}_{i})$

XI. TEST PARA EL EROR DE ESPECIFICACIÓN

El tema no es si α_i es fijo o aleatorio, sino que si la distribución condicional de los α_i dados los Xi es igual a la distribución no condicional de α_i. Así cuando α_i está correlacionado con los Xi, el modelo

$$y_{it} = \mu + \beta' x_{it} + \alpha_i + \varepsilon_{it}$$

Recibe el nombre de modelo de efectos aleatorios.

¿Cómo elegir entre el Modelo de Efectos Fijos y el Modelo de Efectos Aleatorios? Cuando solo se dispone de unas pocas observaciones para diferentes individuos en el tiempo, es excepcionalmente importante hacer el uso más eficiente de los datos a través de los individuos con el fin de estimar aquella parte de las relaciones de comportamiento que contienen las variables que difieren sustancialmente de un individuo a otro. De esta forma, la escasa información temporal disponible será utilizada de manera ventajosa en la estimación de la parte común en la relación económica bajo estudio.

LA PRUEBA DE HAUSMAN (Greene, 1999)

En el procedimiento propuesto por Hausman, bajo la hipótesis nula Ho: $\theta = 0$, es el estimador MCG para el modelo

Si los efectos individuales están correlacionados con las X's, entonces, el modelo de efectos fijos es consistente y eficiente; mientras que el modelo de efectos aleatorios no es consistente.

Si los efectos individuales no están correlacionados con las X's, entonces, el modelo de efectos aleatorios es consistente y eficiente; mientras que el modelo de efectos fijos es consistente pero no eficiente (esto se debería a una subestimación de la variancia).

H_o: El modelo de efectos aleatorios es el modelo correcto.

$$H = (\hat{\beta}_{EA} - \hat{\beta}_{EF})'\left[\Sigma_{EF} - \Sigma_{EA}\right]^{-1}(\hat{\beta}_{EA} - \hat{\beta}_{EF}) \sim \mathrm{X}_k^2 \qquad \textbf{110.}$$

OTRA PRUEBA ALTERNATIVA (Beltrán Barco, 2003)

H_o: La omisión de efectos fijos en el modelo de efectos aleatorios tiene algún efecto en la consistencia de los estimadores de efectos aleatorios ($\gamma = 0$)

$$\tilde{y} = \beta \tilde{x} + \gamma \bar{\bar{x}} + \varepsilon$$

Dónde:

$$\tilde{y} = y_{it} - \theta \bar{y}_i$$
$$\tilde{x} = x_{it} - \theta \bar{x}_i$$
$$\bar{\bar{x}} = x_{it} - \bar{x}_i \qquad \textbf{111.}$$

Entonces se testea que los γ sean cero con un test F en bloques.

¿Cuál de los dos modelos es preferible: el Modelo de Efectos Fijos o el Modelo de Efectos Aleatorios (Beltrán Barco, 2003)?

La respuesta a esta pregunta dependerá del contexto en el cual se encuentre la información y el propósito del análisis. Si la información recabada consiste en una muestra aleatoria de observaciones de una población grande; entonces, el modelo de efectos aleatorios puede ser el más apropiado. Sin embargo, donde no se puede considerar este modelo es cuando se analiza una población completa o grande, en ese caso, es conveniente utilizar el modelo de efectos fijos.

Una de las desventajas potenciales del modelo de efectos aleatorios es que asume que los errores estocásticos asociados con cada unidad observacional están no correlacionados con ningún otro regresor, cuando en muchas circunstancias esto no sucede.

Bibliografía

Castro, J. (2008). *Análisis univariado de series de tiempo.* Notas de Clase BCRP 55º Curso de Extensión Universitaria.

Mayorga M., M., & Muñoz S., E. (2000). *La técnica de datos de panel una guía para su uso e interpretación.* Documento de trabajo del Banco Central de Costa Rica, Banco Central de Costa Rica.

Beltrán Barco, A. (2003). *Econometría de series de tiempo.* Lima: . Universidad del Pacífico. Obtenido de https://econometriaii.files.wordpress.com/2010/01/beltran.pdf

de Arce, R., & Mahía, R. (2007). Técnicas de Previsión de variables financieras:Modelos Arima. (M. d. Citius, Ed.)

Greene, W. (1999). *Análisis Econométrico.* (S. &. Schuster, Trad.) Madrid: Prentice Hall Iberia.

Labra, R. &. (2014). *Guía CERO para datos de panel. Un enfoque práctico.* (U. A. Madrid, Ed.)