ECONOMETRÍA

Iker Caballero Bragagnini

Undergraduate Level

Tabla de contenido

Introducción a la econometría

- La econometría se basa en la aplicación de la estadística matemática y las herramientas de la inferencia estadística para la medición empírica de las relaciones postuladas por la teoría económica

    - Uno de los objetivos principales es hacer un análisis causal: analizar cualitativa y cuantitativamente cómo ciertos factores afectan a una variable asociada a un fenómeno económico de interés

        - Determinar los efectos de ciertas políticas, caracterizar y cuantificar la relación de comportamiento entre variables económicas y simular los efectos de políticas alternativas son algunas de las cosas que permiten los modelos econométricos

    - En econometría se suelen usar datos no experimentales u observacionales, de modo que los valores de la variable objeto de estudio como de los factores susceptibles de afectarla están fuera del control del economista

        - Por eso no se puede interpretar una correlación entre variables como un efecto causal

        - Los datos experimentales, en cambio, se determinan en entornos controlados (se controlan los valores de aquellos factores susceptibles de afectar al fenómeno de estudio

- Dentro de la econometría es posible hacer distinciones entre diversas aplicaciones y áreas

    - La conexión entre los modelos de comportamiento subyacentes y la práctica moderna de la econometría ha incrementado mucho y es útil hacer una distinción entre microeconometría y macroeconometría

        - La microeconometría se caracteriza por el análisis de datos transversales y de panel centrado en los consumidores individuales, las empresas y los decisores a nivel microeconómico. Los practicantes se apoyan en las herramientas teóricas de la microeconomía y los análisis se centran en cuestiones difíciles que suelen requerir formulaciones complejas

        - La macroeconometría se centra en el análisis de series temporales, normalmente de variables agregadas macroeconómicas

- Otra distinción útil es la distinción entre la econometría teórica y la econometría aplicada

    - Los econometristas teóricos desarrollan nuevas técnicas para la estimación y los contrastes de hipótesis y analizan las consecuencias de aplicar métodos particulares cuando las suposiciones que justifican esos métodos no se dan

    - Los econometristas aplicados son los usuarios de estas técnicas y loss analistas de datos reales y simulados

- Los datos económicos se suelen presentar en estructuras variadas, lo cual hace que a veces se tenga que tener en cuenta sus características con tal de poder hacer un análisis econométrico adecuado. Se trabaja principalmente con cuatro tipos de bases de datos: de datos transversales, de secciones transversales agrupadas, de series de tiempo y de datos longitudinales

    - Una base de datos transversales o *cross-sectional data* consiste en una muestra de individuos tomada en un punto en el tiempo

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

        - Una importante característica es que comúnmente se puede asumir que estos datos se han obtenido con muestreo aleatorio de la población, aunque a veces esta suposición no es posible

    - Una base de datos de series temporales consiste en observaciones de una o más variables a lo largo del tiempo. Como el pasado, el presente y el futuro suelen estar conectados, el tiempo es un aspecto importante a tener en cuenta en este tipo de bases de datos

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

- Una dificultad que se encuentra es que no se puede asumir que las observaciones son independientes en el tiempo. Por ello se han desarrollado modificaciones a las técnicas econométricas estándar para tener en cuenta y explotar la naturaleza dependiente de las series de tiempo económicas y para abordar otras cuestiones, como el hecho de que algunas variables económicas tienden a mostrar tendencias claras a lo largo del tiempo

- Otro aspecto importante es la frecuencia con la que se recolectan los datos, dado que muchos datos económicos muestran patrones estacionales que tienen que tenerse en cuenta en el análisis

- Una base de datos de secciones transversales agrupadas es una base de datos que combina datos transversales y series temporales, de modo que se aumenta la muestra con observaciones de diferentes puntos en el tiempo

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57200	16	1100	2	1.5

- De este modo, los individuos de la muestra pueden ser observados en periodos diferentes, pero no se tiene una serie

temporal por cada individuo, si no que un individuo puede ser seleccionado en un punto del tiempo diferente a otro

- Se suele analizar como una base de datos transversales, pero se tiene en cuenta también la dimensión temporal

- Una base de datos de panel o datos longitudinales es una base de datos para la cual se tiene una serie temporal por cada individuo transversal

obsno	city	year	murders	population	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

  - La diferencia entre esta base y la anterior es que para los mismos individuos de la muestra se tienen observaciones de diferentes momentos en el tiempo

- El objetivo principal de hacer un análisis econométrico empírico para comprobar la teoría económica y para evaluar políticas públicas es inferenciar que una variable tiene efecto causal sobre otra

  - La noción de *Ceteris Paribus* (manteniendo todos los otros factores relevantes constantes) juega un papel importante en el análisis causal, dado que la mayoría de cuestiones económicas son por naturaleza *Ceteris Paribus*

    - Si todos los otros factores no se mantienen fijos, no se puede saber el efecto causal de una variable, dado que las otras pueden influir en el resultado

    - Casi nunca es posible mantener absolutamente todos los otros factores constantes, por lo que lo importante suele ser saber si se mantienen fijos suficientes factores como para inferenciar el efecto causal

  - Existen muchas dificultades para poder inferenciar el efecto causal, debido a que se tratan con datos no experimentales u observacionales

El modelo de regresión lineal

- La econometría se centra principalmente en la creación de modelos, por lo tanto, es importante entender cómo se suelen desarrollar estos modelos para poder establecer un marco práctico

    - Un modelo suele comenzar por la observación o proposición de que una variable es causada por o varía junto a otra, o por una proposición cualitativa sobre la relación entre una variable y una o más covariables que se espera que estén relacionadas a la variable interesante en cuestión

        - El modelo suele hacer una afirmación general sobre el comportamiento de los individuos

        - De este modo, un modelo econométrico no nace como un conjunto de ecuaciones, sino como una idea de algún tipo de relación que se tiene que traducir a un conjunto de ecuaciones que permitan responder preguntas importantes de la variable de interés

    - Desde un punto de vista estadístico, se tiene en mente una variable y y un vector de covariables $\boldsymbol{x}$ junto con una distribución de probabilidad conjunta entre ellas $p(y, \boldsymbol{x})$

        - Por lo tanto, la relación entre variables sería el proceso estadístico por el cual se producen las variables

        - Como $\boldsymbol{x}$ es un vector de varias variables aleatorias, se puede descomponer la distribución conjunta en la distribución condicional de y sobre $\boldsymbol{x}$ y en la distribución conjunta de $\boldsymbol{x}$

        $$p(y, \boldsymbol{x}) = p(y|\boldsymbol{x})p(\boldsymbol{x})$$

        - Normalmente uno se interesa en la variación condicional $p(y|\boldsymbol{x})$ de una de las variables en relación a otras más que en la variación conjunta $p(y, \boldsymbol{x})$ de todas las variables

    - En consecuencia, la idea de una distribución conjunta proporciona un buen comienzo para pensar sobre la relación entre la variable de interés y y las covariables $\boldsymbol{x}$

        - Pensando en términos de la distribución condicional, uno se suele centrar en el su valor esperado $E(y|\boldsymbol{x})$, llamado función de regresión

        - Aunque hay otros estadísticos útiles, es necesario establecer como analizar la función de regresión para poder analizar los otros estadísticos

- Aunque existen herramientas más complejas, la herramienta más simple y útil en econometría es el modelo de regresión lineal, dado que suele ser el modelo inicial para la investigación y permite establecer un marco de referencia para las relaciones entre variables

    - El modelo de regresión lineal múltiple se utiliza para estudia la relación entre una variable dependiente o explicada y y una o más variables independientes o explicativas $x_1, x_2, \dots, x_K$, cuya forma genérica es la siguiente:

$$y = f(x_1, x_2, \dots, x_K) + \varepsilon = x_1\beta_1 + x_2\beta_2 + \cdots + x_K\beta_K + \varepsilon$$

        - La teoría económica especifica $f(x_1, x_2, \dots, x_K)$, la cual se conoce como la ecuación de regresión poblacional de y sobre $x_1, x_2, \dots, x_K$

        - Uno de los aspectos más útiles de este modelo es que permite identificar los efectos independientes de un conjunto de variables sobre la variable dependiente

    - En este modelo, y es la variable regresora y $x_1, x_2, \dots, x_K$ son los regresores o las covariables. Estas variables también las suele especificar la teoría económica

        - Para propósitos de modelaje, normalmente es útil pensar en términos de variación autónoma, de modo que se concibe el movimiento de las variables independientes fuera de las relaciones definidas por el modelo mientras que el movimiento de las variables independientes se considera respuesta a algún estímulo independiente o exógeno

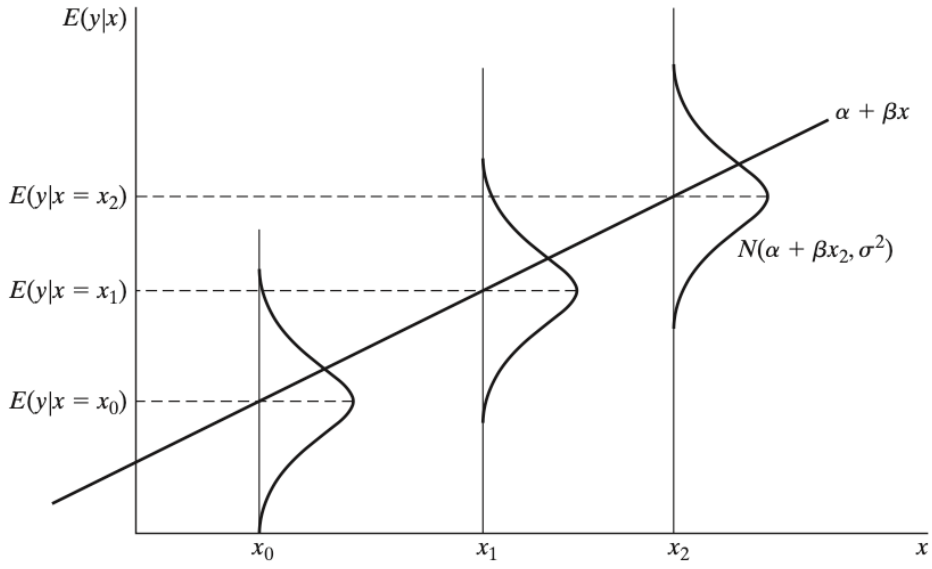
    - El término ε es una perturbación aleatoria o error (porque representa perturbaciones a la relación estable planteada), el cual nace de no poder capturar todas las influencias en una variable económica del modelo

        - El efecto neto de estos factores omitidos puede ser positivo o negativo y se captura en este término de error

        - No obstante, este término de error también se puede debe a errores de medición u otros problemas

- Se asume que cada observación en una muestra $(y_i, x_{i1}, x_{i2}, \dots, x_{i3})$ para $i = 1,2, \dots, n$ se genera por un proceso subyacente descrito por la siguiente ecuación:

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$$



  - El valor observado de y_i es la suma de dos partes: un parte determinística y una parte aleatoria ε_i

  - De este modo, el objetivo es estimar los parámetros desconocidos del modelo, usar los datos para estudiar la validez de las proposiciones teóricas y posiblemente utilizar el modelo parra predecir y

- Siendo el vector columna $\boldsymbol{x_k}$ que contiene las n observaciones en la variable $\boldsymbol{x_k}$ para $k = 1,2, \dots, K$, $\boldsymbol{y}$ el vector columna que contiene las n observaciones $y_1, y_2, \dots, y_K$, $\boldsymbol{\varepsilon}$ el vector columna que contiene las n perturbaciones y $\boldsymbol{\beta}$ el vector columna de parámetros, es posible reformular el modelo de manera más conveniente

  - Uniendo estos datos en una matriz de datos $\boldsymbol{X}$ de tamaño $n \times K$ y asumiendo que la primera columna de $\boldsymbol{X}$ es una columna de unos para que β_1 sea el término constante del modelo, se puede reformular el modelo de regresión lineal múltiple de la siguiente manera:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

  - Para evitar confusiones de notación, se usa $\boldsymbol{x}$ para denotar una columna de $\boldsymbol{X}$, de modo que $\boldsymbol{x_k}$ es la columna k de $\boldsymbol{X}$, y los suscritos j y k se usan para denotar columnas (variables),

mientras que i y t se utilizan para denotar filas (observaciones). Por lo tanto, se puede reescribir el proceso para un individuo:

$$\boldsymbol{x_i'} = (x_{1i}, x_{2i}, \dots, x_{Ki}) \;\; \Rightarrow \;\; \boldsymbol{y_i} = \boldsymbol{x_i'\beta} + \boldsymbol{\varepsilon_i}$$

- En este caso, la matriz $\boldsymbol{X}$ se puede expresar como un vector fila en el que cada componente es $\boldsymbol{x_j}$ (el vector columna de observaciones para la variable j), dado que es equivalente a una matriz adjunta y contiene los mismos datos. También se suele utilizar $\boldsymbol{x}$ para denotar el vector de variables aleatorias

$$\boldsymbol{X} = [\boldsymbol{x_1}|\boldsymbol{x_2}| \dots |\boldsymbol{x_K}] = (\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_K})$$

- El interés principal reside en estimar e inferenciar sobre el vector de parámetros $\boldsymbol{\beta}$

- El modelo de regresión lineal consiste en un conjunto de suposiciones sobre como la base de datos ha sido producida y el proceso generador de datos subyacente

  - Las suposiciones describen la forma del modelo y las relaciones entre las partes, además de sugerir unos procedimientos de estimación e inferencia apropiados. Estas suposiciones son las siguientes:

A1. Linearity: $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$. The model specifies a linear relationship between y and $x_1, \dots, x_K$.

A2. Full rank: There is no exact linear relationship among any of the independent variables in the model. This assumption will be necessary for estimation of the parameters of the model.

A3. Exogeneity of the independent variables: $E[\varepsilon_i \mid x_{j1}, x_{j2}, \dots, x_{jK}] = 0$. This states that the expected value of the disturbance at observation i in the sample is not a function of the independent variables observed at any observation, including this one. This means that the independent variables will not carry useful information for prediction of ε_i.

A4. Homoscedasticity and nonautocorrelation: Each disturbance, ε_i has the same finite variance, σ^2, and is uncorrelated with every other disturbance, ε_j. This assumption limits the generality of the model, and we will want to examine how to relax it in the chapters to follow.

A5. Data generation: The data in $(x_{j1}, x_{j2}, \dots, x_{jK})$ may be any mixture of constants and random variables. The crucial elements for present purposes are the strict mean independence assumption A3 and the implicit variance independence assumption in A4. Analysis will be done conditionally on the observed **X**, so whether the elements in **X** are fixed constants or random draws from a stochastic process will not influence the results. In later, more advanced treatments, we will want to be more specific about the possible relationship between ε_i and $\mathbf{x}_j$.

A6. Normal distribution: The disturbances are normally distributed. Once again, this is a convenience that we will dispense with after some analysis of its implications.

    - Estas son suposiciones que se suelen utilizar en el modelo de regresión lineal, pero no quiere decir que sean todas necesarias para derivar algunos resultados (como el teorema de Gauss-Markov)

    - Las primeras tres suposiciones conforman el modelo de regresión lineal, mientras que las otras especifican las características de las perturbaciones del modelo y las

condiciones para las cuales las observaciones muestrales de $\boldsymbol{x}$ se pueden obtener

- La suposición de linealidad del modelo de regresión incluye un término de error aditivo, de modo que para que un modelo sea lineal en este sentido, es necesario que sea de la forma vista anteriormente con las variables originales o con una transformación adecuada

    - En el contexto de regresión, la linealidad se refiere a la manera en que los parámetros y el término de error entran en la ecuación, y no necesariamente a la relación entre las variables

    - Variaciones de la forma general $f(\boldsymbol{y_i}) = g(\boldsymbol{x_i'\beta} + \boldsymbol{\varepsilon_i})$ permiten crear una gran variedad de formas funcionales que respetan la definición de linealidad

    - Aunque el modelo lineal se puede interpretar como una aproximación para una función subyacente desconocida, es una aproximación muy limitada

    - Aunque el modelo lineal sea muy flexible, no siempre es adecuado para todas las situaciones y no se pueden estimar parámetros ni inferenciar como en este modelo

- La suposición de rango completo o *full rank* establece que no hay relaciones lineales exactas entre variables, también conocida como suposición de no colinearidad perfecta

    - La matriz $\boldsymbol{X}$ es una matriz $n \times K$ con rango K, de modo que $\boldsymbol{X}$ tiene un rango del espacio de columna (y de fila) igual a K y eso quiere decir que las columnas de A son linealmente independientes y que hay como mínimo K observaciones (dado que n debe de ser mayor o igual a K)

    - Esta suposición se conoce como condición de identificación, y es necesaria dado que, de no haber columnas independientes, no se podría estimar los parámetros porque las variables no pueden variar de manera linealmente independiente (variar cuando las otras se quedan constantes) y no se podría inferenciar

    - Hay veces que se pueden utilizar formulaciones que incluyan una misma variable en diferentes formas funcionales (por ejemplo, utilizar cuadrados), pero eso no afecta a la independencia lineal porque aún no siendo funcionalmente independientes, sí que lo son linealmente

- La suposición de exogeneidad de las variables independientes o de independencia de la media expresa que el término de error tiene un valor esperado condicional de cero para cada observación

$$E(\varepsilon_i|\boldsymbol{X}) = 0 \qquad E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \begin{bmatrix} E(\varepsilon_1|\boldsymbol{X}) \\ E(\varepsilon_2|\boldsymbol{X}) \\ \dots \\ E(\varepsilon_n|\boldsymbol{X}) \end{bmatrix} = \mathbf{0}$$

  - En este caso, la media de cada ε_i condicional a todas las observaciones $\boldsymbol{x_i}$ es nula, de modo que ninguna observación en $\boldsymbol{x}$ proporciona información sobre el valor esperado de las perturbaciones o no explica esta variación (ni $\boldsymbol{x_i}$ ni $\boldsymbol{x_j}$)

  - Hay situaciones en las que, aunque $\boldsymbol{x_i}$ no proporcione ninguna información sobre la esperanza condicional del error, $\boldsymbol{x_j}$ puede proporcionar información en otra observación. Este suele ser el caso con series temporales, en donde las observaciones son los periodos y una observación puede de $\boldsymbol{x_j}$ puede proporcionar información sobre el siguiente periodo

  - Esta suposición implica que la esperanza incondicional de la distribución incondicional del término de error también es nula gracias a las propiedades del valor esperado. No obstante, no ocurre lo converso

$$E(\varepsilon_i) = E_{\boldsymbol{X}}[E(\varepsilon_i|\boldsymbol{X})] = E_{\boldsymbol{X}}(0) = 0$$

$$E(\varepsilon_i|\boldsymbol{X}) = 0 \;\Rightarrow\; E(\varepsilon_i) = 0 \quad but\ not \quad E(\varepsilon_i) = 0 \Rightarrow E(\varepsilon_i|\boldsymbol{X}) = 0$$

  - La suposición no suele ser restrictiva, pero asumir $E(\varepsilon_i|\boldsymbol{X}) = 0$ en modelos sin término constante puede alterar el modelo y por eso se asume que los modelos de regresión tienen término constante a no ser que se especifique lo contrario

  - Debido a que para cada ε_i, la covarianza entre $E(\varepsilon_i|\boldsymbol{X})$ y $\boldsymbol{X}$ es lo mismo que $Cov(\varepsilon_i, \boldsymbol{X})$, por lo que la suposición también implica que esta es nula para cualquier i

$$Cov(\varepsilon_i, \boldsymbol{X}), = Cov[E(\varepsilon_i|\boldsymbol{X}), \boldsymbol{X}] =$$

$$= E\left[\Big(E(\varepsilon_i|\boldsymbol{X}) - E_{\boldsymbol{X}}\big(E(\varepsilon_i|\boldsymbol{X})\big)\Big)\big(\boldsymbol{X} - E(\boldsymbol{X})\big)\right] =$$

$$= E\big[\big(E(\varepsilon_i|\boldsymbol{X}) - E_{\boldsymbol{X}}(\varepsilon_i)\big)\big(\boldsymbol{X} - E(\boldsymbol{X})\big)\big] = E\big[0\big(\boldsymbol{X} - E(\boldsymbol{X})\big)\big] = 0$$

- Finalmente, la suposición también implica que la función de regresión $E(\boldsymbol{y}|\boldsymbol{X})$ es $\boldsymbol{X\beta}$, de modo que la regresión de $\boldsymbol{y}$ en $\boldsymbol{X}$ es la esperanza condicional

$$E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X\beta}$$

- La cuarta suposición expresa que la varianza condicional del error con respecto a $\boldsymbol{X}$ es constante y las perturbaciones no están correlacionadas, por lo que la suposición se denomina perturbaciones esféricas

$$Var(\varepsilon_i|\boldsymbol{X}) = \sigma^2 \quad for\ all\ \ i = 1,2,\dots,n$$

$$Cov(\varepsilon_i, \varepsilon_j|\boldsymbol{X}) = 0 \quad for\ all\ \ i \neq j$$

  - La primera parte de la suposición se refiere a que la varianza es constante para todas las observaciones, propiedad conocida como homoscedasticidad. Si la varianza depende de $\boldsymbol{X}$ (o, equivalentemente, varía dependiendo de la observación), entonces la propiedad es la heterocedasticidad

  - La segunda parte de la suposición se refiere a que no hay correlación entre perturbaciones, propiedad conocida como no autocorrelación. Cuando las observaciones siguen una inercia o patrón (hay dependencia unas de otras) se dice que hay autocorrelación

  - No obstante, esta segunda parte no implica que no hay correlación entre y_i y y_j, sino que las desviaciones de las observaciones con respecto a sus valores esperados no están correlacionadas (los términos de error)

  - La no autocorrelación se puede expresar en términos de esperanzas condicionales como la esperanza condicional del producto de las perturbaciones. A partir de esto, es posible expresar la homoscedasticidad en términos de la covarianza y de la esperanza condicional

$$Cov(\varepsilon_i, \varepsilon_j|\boldsymbol{X}) = E\left[\left(\varepsilon_i - E(\varepsilon_i)\right)\left(\varepsilon_j - E(\varepsilon_j)\right)|\boldsymbol{X}\right] =$$

$$= E\left[\left(E(\varepsilon_i|\boldsymbol{X}) - E(\varepsilon_i)\right)\left(E(\varepsilon_j|\boldsymbol{X}) - E(\varepsilon_j)\right)|\boldsymbol{X}\right] = E(\varepsilon_i\varepsilon_j|\boldsymbol{X}) = 0$$

$$for\ all\ \ i \neq j$$

$$Var(\varepsilon_i|\boldsymbol{X}) = Cov(\varepsilon_i, \varepsilon_i|\boldsymbol{X}) = E(\varepsilon_i\varepsilon_i|\boldsymbol{X}) = E(\varepsilon_i^2|\boldsymbol{X}) = \sigma^2$$

$$for\ all\ \ i = 1,2,\dots,n$$

$$Var(\varepsilon_i|\boldsymbol{X}) = E(\varepsilon_i^2|\boldsymbol{X}) - [E(\varepsilon_i|\boldsymbol{X})]^2 = E(\varepsilon_i^2|\boldsymbol{X}) = \sigma^2$$

$$for\ all\ \ i = 1,2, \dots, n$$

- Las dos suposiciones juntas implican que $E(\boldsymbol{\varepsilon\varepsilon}'|\boldsymbol{X}) = \sigma^2\boldsymbol{I}$, lo cual se puede expresar de la siguiente forma:

$$E[\boldsymbol{\varepsilon\varepsilon}'\,|\,\mathbf{X}] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1\,|\,\mathbf{X}] & E[\varepsilon_1\varepsilon_2\,|\,\mathbf{X}] & \cdots & E[\varepsilon_1\varepsilon_n\,|\,\mathbf{X}] \\ E[\varepsilon_2\varepsilon_1\,|\,\mathbf{X}] & E[\varepsilon_2\varepsilon_2\,|\,\mathbf{X}] & \cdots & E[\varepsilon_2\varepsilon_n\,|\,\mathbf{X}] \\ \vdots & \vdots & \vdots & \vdots \\ E[\varepsilon_n\varepsilon_1\,|\,\mathbf{X}] & E[\varepsilon_n\varepsilon_2\,|\,\mathbf{X}] & \cdots & E[\varepsilon_n\varepsilon_n\,|\,\mathbf{X}] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix},$$

- Usando la descomposición de la varianza, se puede ver como la varianza incondicional del término de error es la matriz $\sigma^2\boldsymbol{I}$, lo cual se puede interpretar como la suposición que junta ambas condiciones anteriores (ya que se necesita que no haya otros términos más que la varianza de los errores en la diagonal y eso ocurre cuando no hay correlación entre errores)

$$Var(\boldsymbol{\varepsilon}) = E[Var(\boldsymbol{\varepsilon}|\boldsymbol{X})] + Var[E(\boldsymbol{\varepsilon}|\boldsymbol{X})] = \sigma^2\boldsymbol{I} + 0 = \sigma^2\boldsymbol{I}$$

- La suposición de generación de datos se refiere a que $\boldsymbol{X}$ puede ser fija (no estocástica) o aleatoria (estocástica)

  - Es común suponer que $\boldsymbol{x_i}$ no es estocástica, como en un experimento en donde el analista escoge los valores de los regresores y observa y_i

  - Esta suposición es una conveniencia matemática porque se trataría el vector $\boldsymbol{x_i}$ como una constante en la distribución de probabilidad de $\boldsymbol{x_i}$ y usar métodos estadísticos básicos para obtener los resultados y encontrar propiedades de los estimadores. Además, hace que la tercera y la cuarta suposición sean incondicionales (las esperanzas sean incondicionales)

  - En aplicaciones, lo más común es que los datos en $\boldsymbol{x_i}$ sean aleatorios como y_i. Si $\boldsymbol{x_i}$ es un vector aleatorio, entonces las suposiciones anteriores se transforman en proposiciones sobre la distribución conjunta de y_i y $\boldsymbol{x_i}$

- La naturaleza del regresor y como se entiende el proceso de muestreo son determinantes para la derivación de las propiedades estadísticas de los estimadores y tests de hipótesis

- Sin embargo, $\boldsymbol{X}$ puede contener elementos estocásticos y no estocásticos (variables binarias, constantes, tendencias, etc.) a la vez, por lo que se asume que $\boldsymbol{X}$ puede contener ambos y no solo un tipo

- La suposición de normalidad se refiere que las perturbaciones están normalmente distribuidas con una media nula y con una varianza constante

$$\boldsymbol{\varepsilon} \mid \boldsymbol{X} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

  - En muchos casos, las condiciones del teorema central del límite se darán (aunque sea aproximadamente) y la suposición será razonable

  - Esta suposición implica que las observaciones de las perturbaciones ε_i son estadísticamente independientes y no están correlacionadas

  - Esta suposición no es necesaria para obtener resultados en el modelo de regresión múltiple, pero permite obtener resultados estadísticos más exactos

  - Esta suposición implica que hay independencia estadística entre las perturbaciones, que es un resultado más fuerte que el de independencia de media

Los mínimos cuadrados

- Existen varios métodos para poder estimar los parámetros del modelo, pero el método más popular es el de mínimos cuadrados debido a razones teóricas y prácticas. Este sirve como referencia y otros métodos suelen basarse en modificaciones de este, de modo que es importante estudiar la regresión de mínimos cuadrados

  - Los parámetros desconocidos en la relación estocástica $y_i = \boldsymbol{x_i'\beta} + \varepsilon_i$ son los objetos de estimación, pero es necesario distinguir entre las cantidades poblacionales, como $\boldsymbol{\beta}$ y ε_i, de las estimaciones muestrales, denotadas por $\boldsymbol{b}$ y e_i

    - De este modo, se puede hacer una distinción entre la regresión poblacional $E(y_i|\boldsymbol{x_i}) = \boldsymbol{x_i'\beta}$ y la regresión estimada $\hat{y}_i = \boldsymbol{x_i'b}$

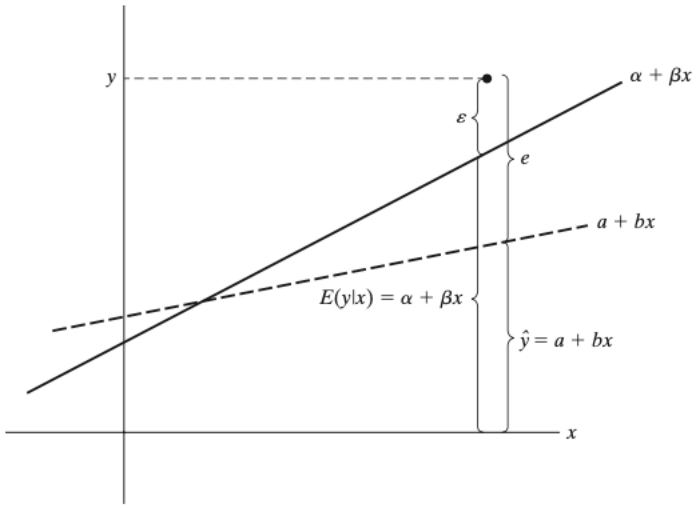
$$E(y_i|\boldsymbol{x_i}) = \boldsymbol{x_i'\beta} \qquad \hat{y}_i = \boldsymbol{x_i'b}$$

- En consecuencia, también se puede hacer una distinción entre la perturbación poblacional asociada a la observación i y la estimación de esta, llamada residuo

$$\varepsilon_i = y_i - \boldsymbol{x_i'\beta} \qquad e_i = y_i - \boldsymbol{x_i'b}$$

- A partir de estas definiciones, se puede ver que se cumplen las siguientes identidades para y_i:

$$y_i = \boldsymbol{x_i'\beta} + \varepsilon_i = \boldsymbol{x_i'b} + e_i$$



- Las cantidades poblacionales $\boldsymbol{\beta}$ es un vector de parámetros desconocidos de la distribución de probabilidad de y_i cuyos valores se intentan estimar a través de los datos $(y_i, \boldsymbol{x_i})$ para $i = 1,2,\dots,n$

- Debido a que se intenta estimar $\boldsymbol{\beta}$ a través de los datos, este es un problema de inferencia estadística, pero se comienza considerando un problema algebraico en el que se tiene que escoger $\boldsymbol{b}$ de modo que $\boldsymbol{x_i'b}$ esté lo más cerca posible a los datos (el criterio de ajuste es la suma de residuos cuadrados), y el candidato más usado es el vector de coeficientes de mínimos cuadrados

  - El vector de coeficientes de mínimos cuadrados minimiza la suma de residuos cuadrados, donde $\boldsymbol{b_0}$ denota la elección del vector de coeficientes

$$\sum_{i=1}^{n} e_{i0}^2 = \sum_{i=1}^{n} (y_i - \boldsymbol{x_i'b_0})^2$$

- En términos matriciales, minimizar la suma de cuadrados requiere escoger $\boldsymbol{b_0}$ tal que se minimice la suma de cuadrados

$$\min_{\boldsymbol{b_0}} S(\boldsymbol{b_0}) = \boldsymbol{e_0'e_0} = (\boldsymbol{y} - \boldsymbol{Xb_0})'(\boldsymbol{y} - \boldsymbol{Xb_0})$$

$$(\boldsymbol{y} - \boldsymbol{Xb_0})'(\boldsymbol{y} - \boldsymbol{Xb_0}) = \boldsymbol{y'y} - \boldsymbol{b_0'X'y} - \boldsymbol{y'Xb_0} - \boldsymbol{b_0'X'Xb_0} =$$

$$= \boldsymbol{y'y} - \boldsymbol{2y'Xb_0} - \boldsymbol{b_0'X'Xb_0}$$

$$\Rightarrow \min_{\boldsymbol{b_0}} S(\boldsymbol{b_0}) = \boldsymbol{e_0'e_0} = \boldsymbol{y'y} - \boldsymbol{2y'Xb_0} - \boldsymbol{b_0'X'Xb_0}$$

- Las condiciones de optimalidad permiten ver como $\boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$, siendo $\boldsymbol{b}$ la solución

$$\frac{\partial S(\boldsymbol{b_0})}{\partial \boldsymbol{b_0}} = -\boldsymbol{2X'y} + \boldsymbol{2X'Xb_0} = \boldsymbol{0} \Rightarrow \boldsymbol{X'Xb_0} = \boldsymbol{X'y}$$

$$\Rightarrow \boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$$

$$\frac{\partial S(\boldsymbol{b_0})}{\partial \boldsymbol{b_0}\partial \boldsymbol{b_0'}} = 2\boldsymbol{X'X} > \boldsymbol{0} \ \ to\ be\ a\ minimum$$

$$\Rightarrow \boldsymbol{c'X'Xc} = \sum_{i=1}^{n} (\boldsymbol{X_i c_i})^{\boldsymbol{2}} > 0 \ \ if\ \boldsymbol{c} \neq \boldsymbol{0} \Rightarrow 2\boldsymbol{X'X} > \boldsymbol{0}$$

- Debido a que $\boldsymbol{X}$ tiene rango K (de modo que $\boldsymbol{Xc} \neq \boldsymbol{0}$ porque no puede ser una combinación lineal de columnas de $\boldsymbol{X}$ igual a $\boldsymbol{0}$), la solución de mínimos cuadrados $\boldsymbol{b}$ es única y minimiza la suma de residuos cuadrados

- A partir de la solución, se puede ver que $\boldsymbol{X'Xb} - \boldsymbol{X'y} = \boldsymbol{0}$, entonces para cada columna $\boldsymbol{x_k}$ de $\boldsymbol{X}$, $\boldsymbol{x_k'e} = 0$. Esto permite desarrollar tres propiedades algebraicas de la solución $\boldsymbol{b}$ siempre que $\boldsymbol{X}$ tenga una columna de unos, denotada por $\boldsymbol{i}$ (si no, no siempre se mantienen)

$$\boldsymbol{X'Xb} - \boldsymbol{X'y} = \boldsymbol{0} \ \Rightarrow \ -\boldsymbol{X'}(\boldsymbol{y} - \boldsymbol{Xb}) = \boldsymbol{0} \ \Rightarrow \ -\boldsymbol{X'e} = \boldsymbol{0} \Rightarrow \boldsymbol{x_k'e} = 0$$

  - La suma de residuos es nula, dado que, si $\boldsymbol{X}$ tiene una columna de unos $\boldsymbol{i}$, entonces $\boldsymbol{i'e} = 0$

$$\boldsymbol{i'e} = \sum_{i=1}^{n} e_i = 0$$

- El hiperplano de la regresión pasa por el punto de medias de los datos, dado que $\boldsymbol{X'Xb - X'y = 0}$ implica que $\bar{y} = \bar{\boldsymbol{x}}'\boldsymbol{b}$

$$-\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{Xb}) = \boldsymbol{0} \;\Rightarrow\; \boldsymbol{y} - \boldsymbol{Xb} = \boldsymbol{0} \;\Rightarrow \boldsymbol{1}'(\boldsymbol{y} - \boldsymbol{Xb}) = \boldsymbol{1}'\boldsymbol{0}$$

$$\Rightarrow \boldsymbol{1}'\boldsymbol{y} - \boldsymbol{1}'\boldsymbol{Xb} = \boldsymbol{0} \;\Rightarrow \left(\frac{1}{n}\right)[\boldsymbol{1}'\boldsymbol{y} - \boldsymbol{1}'\boldsymbol{Xb}] = \left(\frac{1}{n}\right)\boldsymbol{0}$$

$$\Rightarrow \left(\frac{1}{n}\right)\boldsymbol{1}'\boldsymbol{y} - \left(\frac{1}{n}\right)\boldsymbol{1}'\boldsymbol{Xb} = \boldsymbol{0} \;\Rightarrow\; \bar{\boldsymbol{y}} = \bar{\boldsymbol{x}}'\boldsymbol{b}$$

- La media de los valores ajustados de la regresión es equivalente a la media de los valores actuales de $\boldsymbol{y_i}$, dado que $\hat{\boldsymbol{y}} = \boldsymbol{y} + \boldsymbol{e}$

$$\bar{\hat{\boldsymbol{y}}} = \left(\frac{1}{n}\right)\boldsymbol{1}'\hat{\boldsymbol{y}} = \left(\frac{1}{n}\right)\boldsymbol{1}'\boldsymbol{y} + \left(\frac{1}{n}\right)\boldsymbol{1}'\boldsymbol{e} = \left(\frac{1}{n}\right)\boldsymbol{1}'\boldsymbol{y} = \bar{\boldsymbol{y}}$$

- El vector de residuos mínimos cuadrados es $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Xb}$, así que expandiendo $\boldsymbol{b}$ se puede ver que $\boldsymbol{e}$ se puede expresar de la siguiente manera:

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y}$$

  - La matriz $\boldsymbol{M} \equiv \boldsymbol{c}$ es una matriz $n \times n$ es fundamental en el análisis de regresiones y es una matriz simétrica e idempotente ($\boldsymbol{M} = \boldsymbol{M}^2$). Esta se puede interpretar como la matriz que produce el vector de residuos mínimos cuadrados en la regresión de $\boldsymbol{y}$ sobre $\boldsymbol{X}$ cuando se multiplica a la izquierda de $\boldsymbol{y}$ (llamada fabricante de residuos)

  - Cuando se hace una regresión de $\boldsymbol{X}$ sobre $\boldsymbol{X}$, habrá un ajuste perfecto y los residuales serán nulos, de modo que se da la siguiente igualdad:

$$\boldsymbol{MX} = \boldsymbol{0}$$

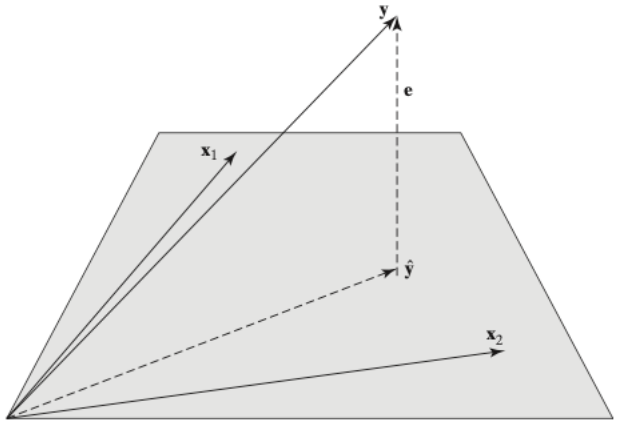
- Como $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Xb}$, se puede ver que $\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{e}$, por lo que los resultados de mínimos cuadrados hagan una partición de $\boldsymbol{y}$ en dos partes: los valores ajustados $\hat{\boldsymbol{y}} = \boldsymbol{Xb}$ y en los residuos $\boldsymbol{e}$, y de este modo se puede derivar la siguiente igualdad:

$$\hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{I} + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y} = (\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}$$

$$where \;\; \boldsymbol{P} \equiv \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

  - Debido a que $\boldsymbol{MX} = \boldsymbol{0}$, $\hat{\boldsymbol{y}}$ y $\boldsymbol{e}$ son ortogonales

- Al igual que $\boldsymbol{M}$, $\boldsymbol{P}$ tiene las propiedades de simetría e idempotencia. Esta matriz $\boldsymbol{P}$ es una matriz de proyección, la cual es formada a partir de $\boldsymbol{X}$ de manera que cuando un vector $\boldsymbol{P}$ se multiplica a la izquierda de $\boldsymbol{y}$, el resultado son los valores ajustados en la regresión de mínimos cuadrados, lo cual es una proyección del vector $\boldsymbol{y}$ en el espacio de columna de $\boldsymbol{X}$ (denotado por W)



- Al multiplicar las matrices $\boldsymbol{M}$ y $\boldsymbol{P}$, se puede ver que estas son ortogonales dada a las propiedades de idempotencia y simetria

$$\boldsymbol{PM} = \boldsymbol{MP} = (\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') =$$

$$= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{0}$$

$$\boldsymbol{PM} = \boldsymbol{MP} = \boldsymbol{0}$$

- Cuando se hace una regresión de $\boldsymbol{X}$ sobre $\boldsymbol{X}$, la proyección será la misma $\boldsymbol{X}$ (de manera lógica)

$$\boldsymbol{PX} = \boldsymbol{X}$$

- Las siguientes identidades son útiles al resolver problemas que involucran resultados de mínimos cuadrados (basadas en la propiedad de idempotencia y simetría de las matrices):

$$\boldsymbol{e}'\boldsymbol{e} = \boldsymbol{y}'\boldsymbol{M}'\boldsymbol{M}\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{M}\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{e} = \boldsymbol{e}'\boldsymbol{y}$$

$$\boldsymbol{e}'\boldsymbol{e} = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{b}$$

- Es común especificar una regresión múltiple cuando solo es interesante un subconjunto del conjunto completo de variables, de modo que es necesario saber que cálculos se involucran para obtener los coeficientes de este subconjunto de variables

- Suponiendo que la regresión involucra dos conjuntos de variables $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$, se puede expresar la ecuación de $\boldsymbol{y}$ de la siguiente manera:

$$\boldsymbol{y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon}$$

  - Las ecuaciones normales permiten obtener una solución para $\boldsymbol{b_1}$ y $\boldsymbol{b_2}$ a través de un sistema de ecuaciones:

$$\begin{bmatrix} \boldsymbol{X_1'X_1} & \boldsymbol{X_1'X_2} \\ \boldsymbol{X_2'X_1} & \boldsymbol{X_2'X_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{b_1} \\ \boldsymbol{b_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_1'y} \\ \boldsymbol{X_2'y} \end{bmatrix} \; as \; \begin{cases} \boldsymbol{X_1'X_1b_1 + X_1'X_2b_2 = X_1'y} \\ \boldsymbol{X_2'X_1b_1 + X_2'X_2b_2 = X_2'y} \end{cases}$$

$$\Rightarrow \boldsymbol{b_1 = (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}X_1'X_2b_2}$$

$$\Rightarrow \boldsymbol{b_2 = (X_2'X_2)^{-1}X_2'y - (X_2'X_2)^{-1}X_2'X_1b_1}$$

  - Como se puede ver, $\boldsymbol{b_1}$ y $\boldsymbol{b_2}$ es el conjunto de coeficientes en la regresión de $\boldsymbol{y}$ en $\boldsymbol{X}$ menos un vector de corrección

- Cuando $\boldsymbol{X_1'X_2 = 0}$ o $\boldsymbol{X_2'X_1 = 0}$, entonces $\boldsymbol{b_1}$ y $\boldsymbol{b_2}$ corresponden a los vectores de coeficientes de las regresiones por separado, lo cual resulta en el siguiente teorema:

  - En la regresión lineal múltiple de mínimos cuadrados de $\boldsymbol{y}$ en dos conjuntos $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$, si los dos conjuntos son ortogonales, entonces los vectores de coeficientes separados se pueden obtener de las regresiones separadas de $\boldsymbol{y}$ sobre $\boldsymbol{X_1}$ e $\boldsymbol{y}$ sobre $\boldsymbol{X_2}$

  - Si los dos conjuntos son ortogonales, entonces $\boldsymbol{X_1'X_2 = X_2'X_1 = 0}$, de modo que $\boldsymbol{b_1 = (X_1'X_1)^{-1}X_1'y}$ y $\boldsymbol{b_2 = (X_2'X_2)^{-1}X_2'y}$, que son los coeficientes que se obtendrían en las regresiones separadas

- Si $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ no son ortogonales, entonces la solución de $\boldsymbol{b_1}$ y $\boldsymbol{b_2}$ se pueden obtener a través del teorema de Frisch-Waugh-Lovell

  - En la regresión lineal de mínimos cuadrados del vector $\boldsymbol{y}$ sobre dos conjuntos de variables $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$, el subvector $\boldsymbol{b_2}$ es el cojunto de coeficientes obtenido cuando se hace una regresiónde los residuos de una regresión de $\boldsymbol{y}$ sobre $\boldsymbol{X_1}$ sobre el conjunto de los residuos obtenidos de la regresión de cada columna de $\boldsymbol{X_2}$ sobre $\boldsymbol{X_1}$

  - A partir de la segunda ecuación del sistema de ecuaciones anterior, se puede sustituir $\boldsymbol{b_1}$ y aislar $\boldsymbol{b_2}$ para poder obtener una expresión de $\boldsymbol{b_2}$ en términos de $\boldsymbol{M_1X_2}$ (matriz de residuos de la regresión de cada columna de $\boldsymbol{X_2}$ sobre $\boldsymbol{X_1}$)

$$\boldsymbol{X_2'X_1b_1 + X_2'X_2b_2 = X_2'y}$$

$$\boldsymbol{X_2'X_1(X_1'X_1)^{-1}X_1'y - X_2'X_1(X_1'X_1)^{-1}X_1'X_2b_2 + X_2'X_2b_2 = X_2'y}$$

$$\boldsymbol{\Rightarrow b_2 = \left[X_2'\left(I - X_1(X_1'X_1)^{-1}X_1'\right)X_2\right]^{-1}\left[X_2'\left(I - X_1(X_1'X_1)^{-1}X_1'\right)y\right]}$$

$$\boldsymbol{= (X_2'M_1X_2)^{-1}(X_2'M_1y)}$$

$$\boldsymbol{\Rightarrow b_2 = \left(X_2^{*\prime}X_2^*\right)^{-1}X_2^*y^* \quad where \quad X_2^* \equiv M_1X_2 \ \& \ y^* \equiv M_1y}$$

- Este proceso normalmente se denomina comúnmente eliminación parcial o compensación del efecto de $\boldsymbol{X_1}$, por lo que los coeficientes en una regresión múltiple normalmente se denominan coeficientes de regresión parcial

- El teorema de Frisch-Waugh-Lovell permite derivar diferentes corolarios que permiten estimar coeficientes de regresores individuales, los coeficientes al incluir un término constante y otros

  - El coeficiente de $\boldsymbol{z}$ en una regresión múltiple de $\boldsymbol{y}$ sobre $\boldsymbol{W = [X|z]}$ (sobre dos grupos $\boldsymbol{X_1 \equiv X}$ y $\boldsymbol{X_2 \equiv z}$, de modo que hay K columnas y una columna adicional $\boldsymbol{z}$) se calcula como $\boldsymbol{c = (z'Mz)^{-1}(z'Mz) = (z^{*\prime}z^*)^{-1}z^*y^*}$ donde $\boldsymbol{M \equiv I - X(X'X)^{-1}X'}$, $\boldsymbol{z^* \equiv Mz}$ e $\boldsymbol{y^* \equiv My}$

    - De este modo, al añadir un regresor al modelo de regresión lineal múltiple, es posible obtener el coeficiente solo con la matriz $\boldsymbol{M}$ y las variables $\boldsymbol{z}$ e $\boldsymbol{y}$, dado que se obtendrán los vectores de residuos de las regresiones descritas anteriormente

  - Los coeficientes en una regresión múltiple que contiene un término constante se pueden obtener transformando los datos a desviaciones de sus medias y haciendo una regresión de la desviación de y sobre las desviaciones de la media muestral de los regresores

    - Considerando el caso en que $\boldsymbol{X_1 \equiv i}$, la solución para $\boldsymbol{b_2}$ en este caso serán los coeficientes de los regresores que no son constantes. El vector de residuos para cualquier variable en $\boldsymbol{X_2}$ será la desviación de la media muestral de esta

$$\boldsymbol{x^* = (I - X_1(X_1'X_1)^{-1}X_1')x = x - i(i'i)^{-1}i'x =}$$

$$\boldsymbol{= x - i\left(\frac{1}{n}\right)i'x = x - i\bar{x} = (I - i\bar{x})x = M^0x}$$

$$\boldsymbol{M^0 \equiv I - i\left(\frac{1}{n}\right)i' = I - i\bar{x}}$$

        - Por lo tanto, cada columna de $\boldsymbol{M_1 X_2}$ es la variable original en forma de desviaciones de sus medias muestrales

    - Si la regresión múltiple de $\boldsymbol{y}$ sobre $\boldsymbol{X}$ contiene un término constante y las variables en la regresión no están correlacionadas, los coeficientes de la regresión múltiple son los mismos que los de las regresiones individuales simples de y sobre la constante y cada variable

        - Cuando el regresor contiene un término constante, se pueden calcular los coeficientes con las regresiones de la desviación de $\boldsymbol{y}$ de su media muestral sobre las columnas de $\boldsymbol{X}$ en forma de desviaciones también

        - De este modo, la ortogonalidad de las columnas significa que las covarianzas muestrales son nulas, debido a que $\boldsymbol{x_k^{*\prime} x_m^*} = Cov(\boldsymbol{x_k}, \boldsymbol{x_m}) = 0$, y eso hace que, si $\boldsymbol{x_k^{*\prime} x_m^*} = 0$, entonces los coeficientes de cada regresor no tengan un vector de corrección (como se ha visto anteriormente)

- El uso de la regresión múltiple hace que se tenga que hacer un experimento conceptual que puede no poder hacerse en la práctica, el análisis *Ceteris Paribus*

    - Para poder realizarlo, se utiliza el coeficiente de correlación parcial, el cual usa el hecho de que cuando se hace una regresión sobre una variable, este no tiene ningún poder explicativo para las perturbaciones (por construcción)

- El criterio de ajuste original (la suma de residuos cuadrados) sugiere que esta es una medida para el ajuste de la línea de regresión a los datos, pero como esta se puede escalar arbitrariamente por un escalar, es mejor analizar el ajuste a partir de si la variación de $\boldsymbol{x}$ es un buen predictor de la variación de y

    - La variación de la variable dependiente se define en términos de la desviación de esta de su media muestral, de modo que la variación total de y es la suma de sus desviaciones cuadradas

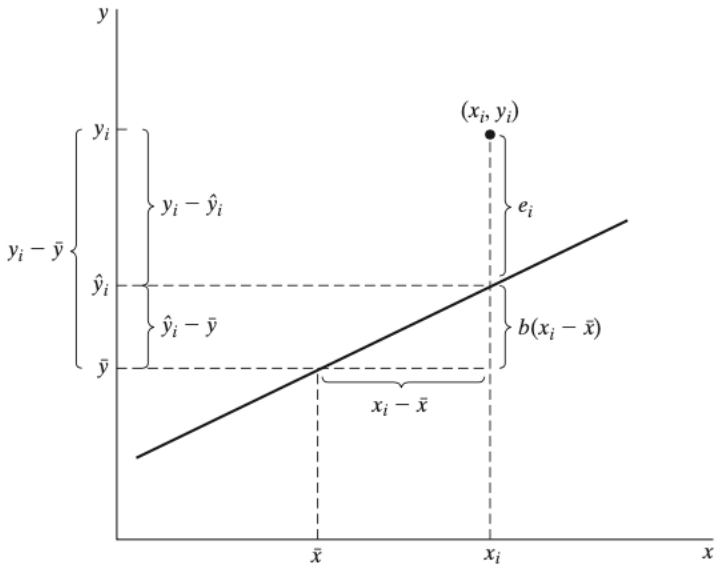
        - En términos de la ecuación de regresión y para una observación individual, se pueden formular las siguientes expresiones:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{e} = \hat{\boldsymbol{y}} + \boldsymbol{e} \qquad y_i = \boldsymbol{x_i'b} + e_i = \hat{y}_i + e_i$$

        - Si la regresión contiene un término constante, entonces las propiedades algebraicas de los mínimos cuadrados anteriormente vistas aplican y se puede ver que $y_i - \hat{y}_i$ se puede expresar de la siguiente manera:

$$y_i - \bar{y}_i = \hat{y}_i - \bar{y}_i + e_i = \boldsymbol{x_i'b} - \boldsymbol{\bar{x}_i'b} + e_i = (\boldsymbol{x_i} - \boldsymbol{\bar{x}_i})'\boldsymbol{b} + e_i$$

- Intuitivamente, la regresión parece que se ajusta mejor si las desviaciones de y de su media muestral se deben más a las desviaciones de $\boldsymbol{x}$ de su media que por los residuos



  - Como ambos términos en la descomposición de $y_i - \bar{y}_i$ suman cero, se utiliza el cuadrado con tal de cuantificar el ajuste. Para el conjunto completo de observaciones, se puede expresar la descomposición en términos de desviaciones de la media muestral

$$\boldsymbol{M^0 y} = \boldsymbol{M^0 X b} + \boldsymbol{M^0 e} = \boldsymbol{M^0 X b} + \boldsymbol{e}$$

  - Como $\boldsymbol{e'M^0X} = \boldsymbol{e'X} = \boldsymbol{0}$, la suma total de cuadrados se puede expresar de la siguiente manera:

$$\boldsymbol{y'M^0 y} = \boldsymbol{b'X'M^0 X b} + \boldsymbol{e'e}$$

$$Total\ Sum\ of\ Sq. = Resid. Sum\ of\ Sq. + Expl. Sum\ of\ Sq.$$

- A partir de esta descomposición, es posible crear una medida de ajuste llamada coeficiente de determinación, denotada por R^2, y que toma valores del cero al uno (incluidos)

$$\frac{ESS}{TSS} = \frac{\boldsymbol{b'X'M^0 X b}}{\boldsymbol{y'M_0 y}} = \boldsymbol{1} - \frac{\boldsymbol{e'e}}{\boldsymbol{y'M_0 y}}$$

  - Este es nulo cuando los coeficientes son nulos (excepto el término constante) y el valor predicho de la regresión es siempre la media de y, de modo que las desviaciones de $\boldsymbol{x}$ de su

media no se traducen en diferentes predicciones de y y $\boldsymbol{x}$ no tiene poder explicativo

- Este es 1 cuando los valores de $\boldsymbol{x}$ e y se encuentran en el mismo hiperplano, de modo que todos los residuos son nulos

- Una manera equivalente de expresar esta medida es con la siguiente ecuación:

$$\hat{\boldsymbol{y}}'\boldsymbol{M_0}\hat{\boldsymbol{y}} = \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{M^0}\boldsymbol{X}\boldsymbol{b} \;\Rightarrow\; \hat{\boldsymbol{y}}'\boldsymbol{M_0}\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}'\boldsymbol{M_0}\boldsymbol{y} \;\Rightarrow R^2 = \frac{\hat{\boldsymbol{y}}'\boldsymbol{M_0}\hat{\boldsymbol{y}}}{\boldsymbol{y}'\boldsymbol{M_0}\boldsymbol{y}}$$

$$\Rightarrow R^2 = \frac{\hat{\boldsymbol{y}}'\boldsymbol{M_0}\hat{\boldsymbol{y}}}{\boldsymbol{y}'\boldsymbol{M_0}\boldsymbol{y}}\frac{\hat{\boldsymbol{y}}'\boldsymbol{M_0}\boldsymbol{y}}{\hat{\boldsymbol{y}}'\boldsymbol{M_0}\hat{\boldsymbol{y}}} \;\Rightarrow\; R^2 = \frac{\sum_{i=1}^{n}\left[(y_i - \bar{y}_i)\left(\hat{y}_i - \bar{\hat{y}}_i\right)\right]^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}_i\right)^2}$$

- El cálculo de R^2 se puede resumir en una tabla de análisis de la varianza como la siguiente:

	Source	*Degrees of Freedom*	*Mean Square*
Regression	$\mathbf{b}'\mathbf{X}'\mathbf{y} - n\bar{y}^2$	$K-1$ (assuming a constant term)	
Residual	$\mathbf{e}'\mathbf{e}$	$n-K$	s^2
Total	$\mathbf{y}'\mathbf{y} - n\bar{y}^2$	$n-1$	$S_{yy}/(n-1) = s_y^2$
Coefficient of determination		$R^2 = 1 - \mathbf{e}'\mathbf{e}/(\mathbf{y}'\mathbf{y} - n\bar{y}^2)$	

- Sin embargo, existen problemas al analizar el ajuste con R^2. La primera dificultad reside en que R^2 siempre aumenta cuanto mayor es el número de regresores

  - Comparando la regresión de $\boldsymbol{y}$ en $\boldsymbol{X}$ con la regresión de $\boldsymbol{y}$ en $\boldsymbol{X}$ y una variable adicional $\boldsymbol{z}$, se asume que la primera produce una suma de residuos cuadrados $\boldsymbol{e}'\boldsymbol{e}$ y la segunda una $\boldsymbol{u}'\boldsymbol{u}$. Además, $c = (\boldsymbol{z}^{*\prime}\boldsymbol{z}^{*})^{-1}(\boldsymbol{z}^{*\prime}\boldsymbol{y}^{*})$ es el coeficiente de la variable $\boldsymbol{z}$, de modo que si se inserta esto en la ecuación del cambio en la suma de errores cuadrados al añadir una variable $\boldsymbol{z}$ se obtiene la siguiente igualdad:

$$\boldsymbol{u}'\boldsymbol{u} = \boldsymbol{e}'\boldsymbol{e} - c^2(\boldsymbol{z}^{*\prime}\boldsymbol{z}^{*}) = \boldsymbol{e}'\boldsymbol{e} - [(\boldsymbol{z}^{*\prime}\boldsymbol{z}^{*})^{-1}(\boldsymbol{z}^{*\prime}\boldsymbol{y}^{*})]^2(\boldsymbol{z}^{*\prime}\boldsymbol{z}^{*}) =$$

$$= \boldsymbol{e}'\boldsymbol{e} - \frac{(\boldsymbol{z}^{*\prime}\boldsymbol{y}^{*})^2}{(\boldsymbol{z}^{*\prime}\boldsymbol{z}^{*})} = \boldsymbol{e}'\boldsymbol{e}\left(1 - r^{*2}_{yz}\right)$$

  - Siendo $R^2_{\boldsymbol{Xz}}$ el coeficiente de determinación en la regresión de $\boldsymbol{y}$ sobre $\boldsymbol{X}$ y la variable adicional $\boldsymbol{z}$ (que, en el caso multivariante, se puede sustituir por un conjunto de variables), $R^2_{\boldsymbol{X}}$ el coeficiente de determinación para la regresión de $\boldsymbol{y}$ sobre $\boldsymbol{X}$ y r^{*2}_{yz} es el coeficiente parcial entre $\boldsymbol{y}$ y $\boldsymbol{z}$ (controlando para $\boldsymbol{X}$), entonces se da la siguiente igualdad:

$$R^2_{\boldsymbol{Xz}} = R^2_{\boldsymbol{X}} + (1 - R^2_{\boldsymbol{X}}) r^{*2}_{yz}$$

- Por lo tanto, el R^2 de una regresión no puede ser menor al añadir un regresor

- El $\bar{R}^2$ es el coeficiente de determinación ajustado (por grados de libertad) se calcula de la siguiente manera:

$$\bar{R}^2 = 1 - \frac{\boldsymbol{e}'\boldsymbol{e}/(n-K)}{\boldsymbol{y}'\boldsymbol{M^0}\boldsymbol{y}/(n-1)}$$

  - Este coeficiente también se puede expresar de manera equivalente en términos de R^2:

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1 - R^2)$$

  - Este coeficiente puede decrecer cuando se añade un regresor adicional y puede ser hasta negativo (de modo que $\bar{R}^2 < 0$). Si $\bar{R}^2$ crece o decrece depende de si la contribución de la nueva variable al ajuste de la regresión más que compensa la corrección por la pérdida adicional de grados de libertad

  - En una regresión múltiple $\bar{R}^2$ decrecerá o crecerá cuando la variable x se elimina de la regresión si el cuadrado de la *t-ratio* asociada a la variable es mayor o menor, respectivamente, a 1

$$t_z^2 = \frac{c^2}{\left(\frac{\boldsymbol{u}'\boldsymbol{u}}{n-K-1}\right)(\boldsymbol{W}'\boldsymbol{W})^{-1}_{K+1,K+1}}$$

- Una segunda dificultad con R^2 está relacionado con el término constante del modelo, dado que si $\boldsymbol{X}$ no tiene una columna de unos, entonces no se puede calcular bien el valor de R^2, dado que pueda dar valores mucho más grandes o pequeños que la misma regresión incluyendo un término constante, puede salir del intervalo $[0,1]$ y hasta podría ser negativo

  - Si no hay una columna de unos en $\boldsymbol{X}$, entonces $\boldsymbol{M^0}\boldsymbol{e} \neq \boldsymbol{e}$ y $\boldsymbol{e}'\boldsymbol{M^0}\boldsymbol{X} \neq 0$ y eso hace que no se pueda eliminar el término $2\boldsymbol{e}'\boldsymbol{M^0}\boldsymbol{X}\boldsymbol{b}$ en $\boldsymbol{y}'\boldsymbol{M^0}\boldsymbol{y} = (\boldsymbol{M^0}\boldsymbol{X}\boldsymbol{b} + \boldsymbol{M^0}\boldsymbol{e})'(\boldsymbol{M^0}\boldsymbol{X}\boldsymbol{b} + \boldsymbol{M^0}\boldsymbol{e})$

- Aunque un criterio que se puede utilizar para poder comparar modelos es el R^2, no hay una referencia para poder comparar modelos porque la bondad del ajuste depende del contexto

- No se puede decir mucho de la calidad relativa de los ajustes de una regresión en diferentes contextos o con diferentes datos, aunque estos provengan del mismo mecanismo generador de datos. Pero, aunque el contexto o los datos sean los mismos, se necesita utilizar una misma referencia a la hora de comparar la bondad de ajuste de los modelos, normalmente relacionado a la variable dependiente

- Además, R^2 es una medida de asociación lineal entre las variables dependientes y la independiente, de modo que, si la relación entre las variables no es lineal, esta será muy baja y puede confundir sobre la bondad del akuste del modelo

- La interpretación de R^2 como la proporción de varianza del modelo explicado por $\boldsymbol{X}$ no es razonable cuando no se utilizan mínimos cuadrados y/o no hay un coeficiente constante

- Es posible estudiar un resultado puramente algebraico que es muy útil para entender el cálculo de modelos de regresión lineal

  - En la regresión de $\boldsymbol{y}$ sobre $\boldsymbol{X}$ se supone que las columnas de $\boldsymbol{X}$ son linealmente transformadas, como multiplicando o dividiendo los datos por una constante (cambio de unidades de medida)

    - En este caso se supone que la matriz $\boldsymbol{X}$ con una transformación lineal es $\boldsymbol{Z}$. De este modo, es interesante saber el efecto de una transformación lineal de la regresión $\boldsymbol{y}$ sobre $\boldsymbol{X}$ comparada con la regresión $\boldsymbol{y}$ sobre $\boldsymbol{Z}$

  - En la regresión lineal de $\boldsymbol{y}$ sobre $\boldsymbol{Z} = \boldsymbol{XP}$ en donde $\boldsymbol{P}$ es una matriz invertible que transforma las columnas de $\boldsymbol{X}$, el vector de coeficientes será $\boldsymbol{P^{-1}b}$ en donde $\boldsymbol{b}$ es el vector de coeficientes de en la regresión lineal de $\boldsymbol{y}$ sobre $\boldsymbol{X}$

    - Los coeficientes de la regresión de $\boldsymbol{y}$ sobre $\boldsymbol{Z}$ serían los siguientes:

$$\boldsymbol{d} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y} = \left((\boldsymbol{XP})'(\boldsymbol{XP})\right)^{-1}(\boldsymbol{XP})'\boldsymbol{y} =$$

$$= (\boldsymbol{P}'\boldsymbol{X}'\boldsymbol{XP})^{-1}\boldsymbol{P}'\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{P}^{-1}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{P}'^{-1}\boldsymbol{P}'\boldsymbol{y} = \boldsymbol{P}^{-1}\boldsymbol{b}$$

    - El vector de residuos $\boldsymbol{u}$ para la regresión de $\boldsymbol{Y}$ sobre $\boldsymbol{Z}$ es idéntico al vector $\boldsymbol{e}$ de la regresión original de $\boldsymbol{Y}$ sobre $\boldsymbol{X}$, de modo que el numerador de $1 - R^2$ es el mismo y el denominador no cambia

$$\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{Z}(\boldsymbol{P^{-1}b}) = \boldsymbol{y} - \boldsymbol{XPP^{-1}b} = \boldsymbol{y} - \boldsymbol{Xb} = \boldsymbol{e}$$

- Este resultado es muy útil porque para casos simples en donde se multipliquen o se dividan las unidades por un número p, el vector de parámetros se multiplica por $1/p$

El estimador de mínimos cuadrados

- Una de las razones por las que el estimador de mínimos cuadrados es tan usado es porque es muy fácil de calcular. Además, es un enfoque natural para la estimación porque utiliza la estructura del modelo y sus suposiciones (solo la de rango completo), es un predictor lineal óptimo para la variable dependiente y porque es el estimador más eficiente bajo unas suposiciones concretas

    - Siendo $\boldsymbol{x}$ un vector de variables independientes en el modelo de regresión poblacional, se asume que los datos pueden ser estocásticos o fijos, que las perturbaciones de la población son ortogonales a las variables independientes (por lo que $E(\boldsymbol{\varepsilon}|\boldsymbol{x}) = 0$ y $Cov(\boldsymbol{\varepsilon}, \boldsymbol{x}) = 0$), entonces se pueden obtener las siguientes identidades:

$$E(\boldsymbol{\varepsilon}|\boldsymbol{x}) = 0 \;\Rightarrow\; E_{\boldsymbol{x}}[E_{\varepsilon}(\boldsymbol{x}\boldsymbol{\varepsilon})] = E_{\boldsymbol{x}}\big[E_{\boldsymbol{y}}\big(\boldsymbol{x}(\boldsymbol{y} - \boldsymbol{x}'\boldsymbol{\beta})\big)\big] = 0$$

$$\Rightarrow\; E_{\boldsymbol{x}}\big[E_{\boldsymbol{y}}(\boldsymbol{x}\boldsymbol{y})\big] = E_{\boldsymbol{x}}(\boldsymbol{x}\boldsymbol{x}')\boldsymbol{\beta}$$

        - Debido a que $\boldsymbol{y} = \boldsymbol{x}'\boldsymbol{\beta}$, $\boldsymbol{x}\boldsymbol{x}'\boldsymbol{\beta}$ no es una función de $\boldsymbol{y}$ y la esperanza se toma para los valores de $\boldsymbol{x}$

        - Utilizando las ecuaciones normales de los mínimos cuadrados, se puede ver que $\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}\boldsymbol{X}'\boldsymbol{b}$ de modo que dividiendo por n se obtiene una expresión en términos de sumatoria, la cual seria una expresión análoga para una muestra:

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i}\, y_i = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i}\, \boldsymbol{x_i'}\right)\boldsymbol{b}$$

        - Si se cumplen las suposiciones de la ley de grandes números, la expresión anterior es un estimador de sus contrapartes anteriores (en términos de esperanzas)

    - Alternativamente, se puede considerar el problema de encontrar un predictor lineal óptimo para $\boldsymbol{y}$. Ignorando suposiciones anteriores sobre la estocasticidad de $\boldsymbol{X}$ y sobre la linealidad del estimador, se puede utilizar la regla del error cuadrático medio (MSE) como función objetivo

        - El valor esperado del error cuadrático medio del predictor se puede expresar de la siguiente manera:

$$MSE = E_y[E_x[(\boldsymbol{y} - \boldsymbol{x}'\boldsymbol{\gamma})^2]] =$$

$$= E_y\left[E_x\left[\left(\boldsymbol{y} - E(\boldsymbol{y}|\boldsymbol{x})\right)^2\right]\right] + E_y\left[E_x[(E(\boldsymbol{y}|\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\gamma})^2]\right]$$

- Por lo tanto, se quiere encontrar el vector $\boldsymbol{\gamma}$ que minimice estas expectativas, por lo que solo es necesario minimizar la segunda parte de la expresión con respecto a $\boldsymbol{\gamma}$ (la otra no depende). Como $E(\boldsymbol{y}|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\gamma}$, el segundo término no depende de $\boldsymbol{y}$ y se pueden poner el operador de derivada dentro de la expresión

$$\frac{\partial E_y\left[E_x[(E(\boldsymbol{y}|\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\gamma})^2]\right]}{\partial \boldsymbol{\gamma}} = E_y\left[E_x\left[\frac{\partial (E(\boldsymbol{y}|\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\gamma})^2}{\partial \boldsymbol{\gamma}}\right]\right] =$$

$$= -2E_y\left[E_x[\boldsymbol{x}(E(\boldsymbol{y}|\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\gamma})]\right] = -2E_y\left[E_x[\boldsymbol{x}(\boldsymbol{x}'\boldsymbol{\gamma} - \boldsymbol{x}'\boldsymbol{\gamma})]\right] = 0$$

$$\Rightarrow\ E_y\left[E_x[\boldsymbol{x}E(\boldsymbol{y}|\boldsymbol{x})]\right] = E_x\left[E_y(\boldsymbol{x}\boldsymbol{x}')\right]\boldsymbol{\gamma}$$

- El resultado anterior permite obtener la condición necesaria para obtener el mínimo del MSE, la cual es la misma equivalencia encontrada anteriormente:

$$E_y\left[E_x[\boldsymbol{x}(E(\boldsymbol{y}|\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\gamma})]\right] = Cov(\boldsymbol{x}, \boldsymbol{y}) + E(\boldsymbol{x})E_x[E(\boldsymbol{y}|\boldsymbol{x})] =$$

$$= Cov(\boldsymbol{x}, \boldsymbol{y}) + E(\boldsymbol{x})E(\boldsymbol{y}) = E_x\left[E_y(\boldsymbol{x}y)\right]$$

$$\Rightarrow\ E_x\left[E_y(\boldsymbol{x}y)\right] = E_x\left[E_y(\boldsymbol{x}\boldsymbol{x}')\right]\boldsymbol{\gamma}$$

  - Si el mecanismo de generación de datos $(x_i, y_i)_{i=1,\dots,n}$ es tal que la ley de grandes números aplica a los estimadores $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i}\, y_i = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i}\, \boldsymbol{x_i'}\right)\boldsymbol{b}$ para las matrices $\boldsymbol{X}'\boldsymbol{y}$ y $\boldsymbol{X}\boldsymbol{X}'\boldsymbol{b}$, entonces el predictor lineal mínimo del error cuadrático esperado de y_i es la línea de regresión por mínimos cuadrados

    - Asumiendo que los valores esperados existen, se pueden estimar con las sumatorias anteriormente vistas y esto significa que, sin importar la forma de la media condicional (la especificación de la regresión), el estimador de mínimos cuadrados es el predictor lineal que minimiza el MSE

- El estimador de mínimos cuadrados ordinario cumple varias propiedades, entre las cuales hay la de no estar sesgado o su eficiencia bajo las suposiciones del teorema de Gauss-Markov

- El estimador de mínimos cuadrados ordinario no tiene sesgo en ninguna de las muestras. Esto se puede demostrar haciendo una expansión del vector de parámetros y de $\boldsymbol{y}$ y tomando el valor esperado condicional a $\boldsymbol{X}$

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

$$\Rightarrow E(\boldsymbol{b}|\boldsymbol{X}) = E(\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{\beta} + E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}|\boldsymbol{X}) =$$

$$= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{\beta}$$

  - Consecuentemente, se obtiene que la esperanza de $\boldsymbol{b}$ es necesariamente $\boldsymbol{\beta}$

$$E(\boldsymbol{b}) = E_{\boldsymbol{X}}[E(\boldsymbol{b}|\boldsymbol{X})] = E_{\boldsymbol{X}}[\boldsymbol{\beta}] = \boldsymbol{\beta}$$

  - Por lo tanto, para cualquier conjunto particular de observaciones $\boldsymbol{X}$, el estimador de mínimos cuadrados ordinario no tiene sesgo (su esperanza es el vector de parámetros poblacionales)

$$E(\boldsymbol{b}) - \boldsymbol{\beta} = \boldsymbol{\beta} - \boldsymbol{\beta} = 0$$

  - Se ha condicionado en $\boldsymbol{X}$ y no en $\boldsymbol{x_i}$ porque el estimador de mínimos cuadrados se puede entender como la contraparte muestral al vector de parámetros del predictor mínimo del MSE $\boldsymbol{\gamma}$, el cual pertenece la función poblacional, y porque, desde un punto de vista bayesiano, se puede entender que $E(\boldsymbol{b}|\boldsymbol{X}) = \boldsymbol{\beta}$ se espera que el estimador se comporte como $\boldsymbol{\beta}$ dados los datos muestrales finitos $\boldsymbol{X}$

- En todo momento se ha utilizado la suposición de que la especificación correcta del modelo de regresión es $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$, pero hay muchos tipos de errores de especificación que se pueden hacer al construir el modelo de regresión. Los más comunes son la omisión de variables relevantes y la inclusión de variables superfluas

  - Suponiendo que el modelo correctamente especificado es un modelo de la forma $\boldsymbol{y} = \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{X_2}\boldsymbol{\beta_2} + \boldsymbol{\varepsilon}$ (en donde $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ son una partición de $\boldsymbol{X}$ con K_1 y K_2 columnas respectivamente), si se hace una regresión de $\boldsymbol{y}$ sobre $\boldsymbol{X_1}$ (sin incluir $\boldsymbol{X_2}$), el estimador y la esperanza condicionada de este a la muestra entera $\boldsymbol{X}$ permite obtener la fórmula de variable omitida

$$\boldsymbol{b_1} = (\boldsymbol{X_1'}\boldsymbol{X_1})^{-1}\boldsymbol{X_1'}\boldsymbol{y} = (\boldsymbol{X_1'}\boldsymbol{X_1})^{-1}\boldsymbol{X_1'}(\boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{X_2}\boldsymbol{\beta_2} + \boldsymbol{\varepsilon}) =$$

$$= (\boldsymbol{X_1'}\boldsymbol{X_1})^{-1}\boldsymbol{X_1'}\boldsymbol{X_1}\boldsymbol{\beta_1} + (\boldsymbol{X_1'}\boldsymbol{X_1})^{-1}\boldsymbol{X_1'}\boldsymbol{X_2}\boldsymbol{\beta_2} + (\boldsymbol{X_1'}\boldsymbol{X_1})^{-1}\boldsymbol{X_1'}\boldsymbol{\varepsilon} =$$

$$= \boldsymbol{\beta_1} + (\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'X_2\beta_2} + (\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'\varepsilon}$$

$$E(\boldsymbol{b_1}|\boldsymbol{X}) = \boldsymbol{\beta_1} + (\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'X_2\beta_2} + (\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'}E(\boldsymbol{c}|\boldsymbol{X}) =$$

$$= \boldsymbol{\beta_1} + (\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'X_2\beta_2} = \boldsymbol{\beta_1} + \boldsymbol{P_{1,2}\beta_2}$$

$$where \quad \boldsymbol{P_{1,2}} \equiv (\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'X_2}$$

- La matriz $\boldsymbol{P_{1,2}}$ es una matriz $K_1 \times K_2$ la cual representa una matriz cuyas columnas son los parámetros de las variables para la regresión de las columnas de $\boldsymbol{X_2}$ sobre las de $\boldsymbol{X_1}$

- Con la fórmula de la omisión de variables, es fácil deducir la dirección del sesgo en casos simples como el caso univariante en el que se omite una variable, pero si en $\boldsymbol{X_1}$ y en $\boldsymbol{X_2}$ hay más variables, entonces los múltiples coeficientes pueden tener diferentes signos y la dirección no es clara. En el caso simple, $\boldsymbol{P_{1,2}}$ se puede entender como la *ratio* entre la covarianza de $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ y la varianza $\boldsymbol{X_1}$, pero como la esperanza está condicionada a $\boldsymbol{X}$, el resultado es constante

$$\frac{Cov(\boldsymbol{X_1}, \boldsymbol{X_2})}{Var(\boldsymbol{X_1})} = E[(\boldsymbol{X_1'X_1})^{-1}|\boldsymbol{X}]E[\boldsymbol{X_1'X_2}|\boldsymbol{X}] = (\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'X_2}$$

- Si la especificación correcta de la regresión es $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ y se estima como si $\boldsymbol{y} = \boldsymbol{X_1\beta_1} + \boldsymbol{X_2\beta_2} + \boldsymbol{\varepsilon}$ fuera correcta, entonces se estarían incluyendo variables innecesarias. No obstante, el modelo no es incorrecto, sino que se estima $\boldsymbol{\beta}$ correctamente, pero se falla en imponer la restricción $\boldsymbol{\beta_2} = \boldsymbol{0}$ (en el caso anterior, se estimaba $\boldsymbol{\beta}$ sujeto a una restricción incorrecta $\boldsymbol{\beta_2} = \boldsymbol{0}$)

$$E(\boldsymbol{b}|\boldsymbol{X}) = \begin{pmatrix}\boldsymbol{\beta_1}\\ \boldsymbol{\beta_2}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\beta_1}\\ \boldsymbol{0}\end{pmatrix}$$

- El problema de esto reside en que incluir variables irrelevantes hace que la precisión de los estimadores, en el sentido que la matriz de varianzas y covarianzas de la regresión de $\boldsymbol{y}$ sobre $\boldsymbol{X_1}$ es siempre menor a la de la regresión de $\boldsymbol{y}$ sobre $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$

- Si los regresores se pueden tratar como valores fijos, entonces la varianza muestral del estimador de mínimos cuadrados se puede derivar tratando $\boldsymbol{X}$ como una matriz de constantes. Alternativamente, se puede permitir que $\boldsymbol{X}$ sea estocástica y condicionar la esperanza a $\boldsymbol{X}$

- Considerando el vector de parámetros, si se define $\boldsymbol{A} \equiv (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, entonces bajo las suposiciones hechas, $\boldsymbol{b}$ es un estimador lineal y no tiene sesgo independientemente de la distribución de $\boldsymbol{\varepsilon}$ (solo se necesita la suposición de independencia de media)

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

$$E(\boldsymbol{b}|\boldsymbol{X}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{\beta}$$

- A partir de la suposición de que la varianza es constante y de la simetría de $\boldsymbol{X}'\boldsymbol{X}$, se puede obtener la varianza del estimador de mínimos cuadrados

$$Var(\boldsymbol{b}|\boldsymbol{X}) = E[(\boldsymbol{b} - \boldsymbol{\beta})(\boldsymbol{b} - \boldsymbol{\beta})'|\boldsymbol{X}] =$$

$$= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon})'|\boldsymbol{X}] =$$

$$= E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1'}\middle|\boldsymbol{X}\right] = E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}]$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} =$$

$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})}$$

- En el modelo de regresión lineal con una matriz de regresores $\boldsymbol{X}$, el estimador de mínimos cuadrados $\boldsymbol{b}$ es el estimador lineal sin sesgo de mínima varianza (cc) de $\boldsymbol{\beta}$. Para cualquier vector de constantes $\boldsymbol{w}$, el BLUE de $\boldsymbol{w}'\boldsymbol{\beta}$ es $\boldsymbol{w}'\boldsymbol{b}$, donde $\boldsymbol{b}$ es el estimador de mínimos cuadrados

  - El teorema no utiliza la suposición de que las perturbaciones se distribuyen de manera normal, por lo que solo se necesitan las primeras cuatro (la quinta se puede usar o no)

  - Para poder demostrar este teorema, se puede proponer el uso de un estimador lineal $\boldsymbol{b}_L = \boldsymbol{C}\boldsymbol{y}$ tal que $E[\boldsymbol{b}_L|\boldsymbol{X}] = \boldsymbol{\beta}$ y encontrar un estimador de esta clase que tenga una matriz de varianzas y covarianzas menor a $\boldsymbol{b}$. Esto implica que $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{I}$ y, por tanto, hay varios candidatos para $\boldsymbol{C}$

$$E(\boldsymbol{b}_L|\boldsymbol{X}) = E(\boldsymbol{C}\boldsymbol{y}|\boldsymbol{X}) = E(\boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{\beta} \;\Rightarrow\; \boldsymbol{C}\boldsymbol{X} = \boldsymbol{I}$$

  - Definiendo una matriz $\boldsymbol{D} \equiv \boldsymbol{C} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, es posible demostrar que la varianza del estimador $\boldsymbol{b}_L$ es mayor a la de $\boldsymbol{b}$ porque $\boldsymbol{q}'\boldsymbol{D}\boldsymbol{D}'\boldsymbol{q} \geq 0$ (la matriz es semidefinida positiva), obteniendo así el siguiente resultado:

$$Var(\boldsymbol{b_L}|\boldsymbol{X}) = Var(\boldsymbol{b}|\boldsymbol{X}) + \sigma^2\boldsymbol{DD}' = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} + \sigma^2\boldsymbol{DD}'$$

- La demostración en el caso de $\boldsymbol{w}'\boldsymbol{b}$ seguiría un patrón igual al de esta demostración y se comprobaría la segunda proposición

- Si se permite que los regresores sean estocásticos, entonces se pueden encontrar resultados iguales no condicionados a los regresores $\boldsymbol{X}$ a través de propiedades del valor esperado

  - Para obtener la varianza incondicional de $\boldsymbol{b}$, se puede utilizar la ley de esperanzas iteradas y el hecho de que $E(\boldsymbol{b}|\boldsymbol{X}) = \boldsymbol{\beta}$

$$Var(\boldsymbol{b}) = E_{\boldsymbol{X}}[Var(\boldsymbol{b}|\boldsymbol{X})] + Var_{\boldsymbol{X}}[E(\boldsymbol{b}|\boldsymbol{X})] =$$

$$= E_{\boldsymbol{X}}[\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}] + Var_{\boldsymbol{X}}(\boldsymbol{\beta}) = \sigma^2 E_{\boldsymbol{X}}[(\boldsymbol{X}'\boldsymbol{X})^{-1}]$$

  - Este resultado permite ver que la varianza de $\boldsymbol{b}$ solo se puede describir en términos del comportamiento medio de $\boldsymbol{X}$. Además, utilizando la demostración anterior, se puede ver como $Var(\boldsymbol{b}|\boldsymbol{X}) \leq Var(\boldsymbol{b_L}|\boldsymbol{X})$, pero debido a que esto se cumple para cualquier $\boldsymbol{X}$ particular, también se debe cumplir para $Var(\boldsymbol{b}) = E_{\boldsymbol{X}}[Var(\boldsymbol{b}|\boldsymbol{X})]$, por lo que se da la siguiente desigualdad:

$$Var(\boldsymbol{b}|\boldsymbol{X}) \leq Var(\boldsymbol{b_L}|\boldsymbol{X}) \;\Rightarrow\; Var(\boldsymbol{b}) \leq Var(\boldsymbol{b_L})$$

  - En consecuencia, en el modelo de regresión lineal, el estimador de mínimos cuadrados ordinarios es un MVLUE independientemente de si $\boldsymbol{X}$ es estocástico o no mientras se cumplan las otras suposiciones

- Para poder hacer contrastes de hipótesis sobre $\boldsymbol{\beta}$ o para construir intervalos de confianza, se requiere un estimado para la matriz de varianzas y covarianzas. Una manera natural de estimar $Var(\boldsymbol{b}|\boldsymbol{X})$ es a través de los residuos, debido a que las perturbaciones cumplen $Var(\varepsilon_i|\boldsymbol{X}) = E({\varepsilon_i}^2|\boldsymbol{X}) - E(\varepsilon_i|\boldsymbol{X})^2 = E({\varepsilon_i}^2|\boldsymbol{X}) = \sigma^2$ y los residuos son estimadores de los errores

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} {e_i}^2$$

  - No obstante, los residuos son unos estimadores imperfectos de las perturbaciones porque e_i depende de $\boldsymbol{\beta}$ y esta no se puede observar directamente

$$e_i = y_i - \boldsymbol{x_i'}\boldsymbol{b} = \boldsymbol{x_i'}\boldsymbol{\beta} + \varepsilon_i - \boldsymbol{x_i'}\boldsymbol{b} = \varepsilon_i + \boldsymbol{x_i'}(\boldsymbol{b} - \boldsymbol{\beta})$$

- Formalmente, se puede derivar $\sum_{i=1}^{n} {e_i}^2$ para poder ver que este es igual a $\sigma^2(n-K)$

$$E(\boldsymbol{e}'\boldsymbol{e}|\boldsymbol{X}) = E((\boldsymbol{M}\boldsymbol{\varepsilon})'\boldsymbol{M}\boldsymbol{\varepsilon}|\boldsymbol{X}) = E(\boldsymbol{\varepsilon}'\boldsymbol{M}'\boldsymbol{M}\boldsymbol{\varepsilon}|\boldsymbol{X}) = E(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}|\boldsymbol{X}) =$$

$$= E(tr(\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon})|\boldsymbol{X}) = E(tr(\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')|\boldsymbol{X}) = E(\boldsymbol{M}tr(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')|\boldsymbol{X}) =$$

$$= tr[\boldsymbol{M}E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X})] = tr(\boldsymbol{M}\sigma^2\boldsymbol{I}) = \sigma^2 tr(\boldsymbol{M}) = \sigma^2(n-K)$$

$$as\ \ tr(\boldsymbol{M}) = tr(\boldsymbol{I_n}) - tr(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') =$$

$$= tr(\boldsymbol{I_n}) - tr(\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) = tr(\boldsymbol{I_n}) - tr(\boldsymbol{I_K}) = n-K$$

- Este resultado muestra que el estimador está sesgado hacia cero (aunque se vuelva menor cuando el tamaño de la muestra n incrementa). Un estimador no sesgado (tanto condicionalmente como de manera incondicional) sería $s^2 \equiv \boldsymbol{e}'\boldsymbol{e}/(n-K)$, y el error estándar de la regresión sería s

$$s^2 \equiv \frac{\boldsymbol{e}'\boldsymbol{e}}{n-K}$$

- Por lo tanto, con s^2 se puede hacer una estimación de la varianza de $\boldsymbol{b}$ condicional a $\boldsymbol{X}$ con la muestra, y a su vez, obtener el error estándar para cada b_k a través de la raíz del elemento kk (de la diagonal) de la matriz

$$Est.Var(\boldsymbol{b}|\boldsymbol{X}) = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

- Hasta ahora no se ha usado la suposición de normalidad de las perturbaciones, la cual es una suposición bastante útil para poder construir intervalos de confianza y tests de hipótesis

  - Si se asume que las perturbaciones siguen una distribución normal multivariante, entonces se puede comprobar la distribución del vector de parámetros y de los estimadores siguen una distribución normal multivariante

$$\boldsymbol{b} \mid \boldsymbol{X} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}) \qquad b_k \mid \boldsymbol{X} \sim N\left(\beta_k, \sigma^2(\boldsymbol{X}'\boldsymbol{X})_{kk}^{-1}\right)$$

  - Si no se hace suposiciones sobre la distribución de las perturbaciones, no es posible decir nada de la distribución de parámetros (ya sea condicional o incondicional)

- Hasta ahora se han establecido resultados para una muestra finita, pero también es posible describir las propiedades asintóticas o de muestra grande del estimador de mínimos cuadrados

    - Aunque se haya visto que el estimador no es sesgado, basarse solo en un criterio de sesgo no es muy útil

        - Esto es debido a que en la mayoría de casos, los estimadores estarán sesgados y solo se puede esperar que estos tiendan a no tener sesgo cuanta más información se use, y, además, que un estimador no esté sesgado no implica que más información sea mejor para la estimación de los parámetros. La propiedad de consistencia mejora el criterio del sesgo en ambas direcciones

        - Siendo $\boldsymbol{X}$ estocástica o fija, se asume que $(\boldsymbol{x_i}, \boldsymbol{\varepsilon_i})$ es una secuencia de observaciones independientes y se asume que $(1/n)\boldsymbol{X}\boldsymbol{X}' \xrightarrow{P} \boldsymbol{Q}$ cuando $n \to \infty$, donde $\boldsymbol{Q}$ es una matriz semidefinida positiva. Esto permite obtener los siguientes resultados:

$$\boldsymbol{b} = \boldsymbol{\beta} + \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{\varepsilon}\right)$$

$$\operatorname*{plim}_{n\to\infty} \boldsymbol{b} = \boldsymbol{\beta} + \boldsymbol{Q}^{-1}\left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{X}'\boldsymbol{\varepsilon}\right) = \boldsymbol{\beta} + \boldsymbol{Q}^{-1}\left(\operatorname*{plim}_{n\to\infty}\bar{\boldsymbol{w}}\right)$$

$$where \quad \boldsymbol{w_i} \equiv \boldsymbol{x_i}\varepsilon_i \quad \& \quad \bar{\boldsymbol{w}} \equiv \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{w_i} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x_i}\varepsilon_i = \frac{1}{n}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

        - A partir de las suposiciones, es posible obtener la esperanza y la varianza (condicional e incondicional) del estimador para poder comprobar como el estimador es consistente con las suposiciones hechas

$$E(\boldsymbol{w_i}) = E_{\boldsymbol{X}}[E(\boldsymbol{w_i}|\boldsymbol{x_i})] = E_{\boldsymbol{X}}[\boldsymbol{x_i}E(\varepsilon_i|\boldsymbol{x_i})] = 0$$

$$\Rightarrow E(\bar{\boldsymbol{w}}) = \frac{1}{n}E\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{w_i}\right) = 0$$

$$\Rightarrow E(\bar{\boldsymbol{w}}|\boldsymbol{X}) = \frac{1}{n}\boldsymbol{X}'E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = 0$$

$$Var(\bar{\boldsymbol{w}}|\boldsymbol{X}) = E(\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}'|\boldsymbol{X}) = E\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}\frac{1}{n}\middle|\boldsymbol{X}\right) =$$

$$= \frac{1}{n}\boldsymbol{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X})\boldsymbol{X}\frac{1}{n} = \frac{1}{n}\boldsymbol{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X})\boldsymbol{X}\frac{1}{n} = \frac{1}{n}\boldsymbol{X}'\sigma^2\boldsymbol{I}\boldsymbol{X}\frac{1}{n} =$$

$$= \left(\frac{\sigma^2}{n}\right)\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)$$

$$Var(\bar{\boldsymbol{w}}) = E_{\boldsymbol{X}}[Var(\bar{\boldsymbol{w}}|\boldsymbol{X})] + Var_{\boldsymbol{X}}[E(\bar{\boldsymbol{w}}|\boldsymbol{X})] = \left(\frac{\sigma^2}{n}\right)E_{\boldsymbol{X}}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)$$

- Debido a la suposición hecha sobre la convergencia de $(1/n)\boldsymbol{X}'\boldsymbol{X}$, tanto la esperanza como la varianza son nulas cuando $n \to \infty$ y eso hace que la media cuadrática tienda a cero y que, por tanto, $(1/n)\boldsymbol{X}'\boldsymbol{\varepsilon} \xrightarrow{P} 0$, haciendo que $\boldsymbol{b}$ sea un estimador consistente

$$\lim_{n\to\infty}\left(\frac{\sigma^2}{n}\right)\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right) = 0\boldsymbol{Q} = \boldsymbol{0}$$

$$\Rightarrow \left(\frac{1}{n}\right)\boldsymbol{X}'\boldsymbol{\varepsilon} \xrightarrow{P} 0 \;\Rightarrow\; \operatorname*{plim}_{n\to\infty}\boldsymbol{b} = \boldsymbol{\beta} + \boldsymbol{Q}^{-1}\,\boldsymbol{0} = \boldsymbol{\beta}$$

- Las suposiciones hechas pueden ser muy restrictivas dependiendo del contexto, por lo que las condiciones de Grenander suelen ser más débiles y se cumplen en casi todas las ocasiones

G1. For each column of $\mathbf{X}$, $\mathbf{x}_k$, if $d_{nk}^2 = \mathbf{x}_k'\mathbf{x}_k$, then $\lim_{n\to\infty} d_{nk}^2 = +\infty$. Hence, $\mathbf{x}_k$ does not degenerate to a sequence of zeros. Sums of squares will continue to grow as the sample size increases. No variable will degenerate to a sequence of zeros.

G2. $\text{Lim}_{n\to\infty} x_{ik}^2/d_{nk}^2 = 0$ for all $i = 1, \ldots, n$. This condition implies that no single observation will ever dominate $\mathbf{x}_k'\mathbf{x}_k$, and as $n \to \infty$, individual observations will become less important.

G3. Let $\mathbf{R}_n$ be the sample correlation matrix of the columns of $\mathbf{X}$, excluding the constant term if there is one. Then $\lim_{n\to\infty} \mathbf{R}_n = \mathbf{C}$, a positive definite matrix. This condition implies that the full rank condition will always be met. We have already assumed that $\mathbf{X}$ has full rank in a finite sample, so this assumption ensures that the condition will never be violated.

- Otra suposición de la que se puede prescindir es la suposición de normalidad de las perturbaciones. Bajo suposiciones generales razonables sobre los procesos generadores de datos muestrales, las distribuciones asintóticas proporcionan una base para la inferencia estadística en el modelo de regresión

  - Usando el teorema central del límite y suponiendo que las observaciones son independientes, se puede obtener el tipo de resultado anteriormente visto

$$\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1} \xrightarrow{P} \boldsymbol{Q}^{-1} \;\;\&\;\; \left(\frac{1}{\sqrt{n}}\right)\boldsymbol{X}'\boldsymbol{\varepsilon} \xrightarrow{D} N(\boldsymbol{0}, \sigma^2\boldsymbol{Q})$$

$$\Rightarrow \quad \boldsymbol{Q}^{-1}\left(\frac{1}{\sqrt{n}}\right)\boldsymbol{X}'\boldsymbol{\varepsilon} \xrightarrow{D} N\left(\boldsymbol{Q}^{-1}\mathbf{0}, \boldsymbol{Q}^{-1}(\sigma^2\boldsymbol{Q})\boldsymbol{Q}^{-1'}\right)$$

$$\Rightarrow \quad \boldsymbol{Q}^{-1}\left(\frac{1}{\sqrt{n}}\right)\boldsymbol{X}'\boldsymbol{\varepsilon} \xrightarrow{D} N(\mathbf{0}, \sigma^2\boldsymbol{Q}^{-1})$$

$$\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta}) = \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\right)\boldsymbol{X}'\boldsymbol{\varepsilon} \;\Rightarrow\; \sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta}) \xrightarrow{D} N(\mathbf{0}, \sigma^2\boldsymbol{Q}^{-1})$$

- Si $\boldsymbol{\varepsilon}$ se distribuyen independientemente con media nula y varianza finita σ^2, y $\boldsymbol{X}$ es tal que las condiciones de Grenander se cumplen, entonces la distribución asintótica de $\boldsymbol{b}$ es la siguiente:

$$\boldsymbol{b} \approx N\left(\boldsymbol{\beta}, \frac{\sigma^2}{n}\boldsymbol{Q}^{-1}\right)$$

- Para poder completar la derivación de las propiedades asintóticas de $\boldsymbol{b}$ es necesario estimar la varianza asintótica $(\sigma^2/n)\boldsymbol{Q}^{-1}$ a través de s^2

$$s^2 = \frac{1}{n-K}\boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}$$

  - Expandiendo el término, se obtiene una expresión con la cual se puede comprobar que el estimador s^2 tiende a la media de los errores cuadráticos

$$s^2 = \frac{1}{n-K}\boldsymbol{\varepsilon}'(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{\varepsilon} =$$

$$= \frac{1}{n-K}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}) =$$

$$= \frac{n}{n-K}\left[\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n} - \left(\frac{\boldsymbol{\varepsilon}'\boldsymbol{X}}{n}\right)\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{n}\right)\right]$$

$$\Rightarrow \frac{n}{n-K}\left[\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n} - \left(\frac{\boldsymbol{\varepsilon}'\boldsymbol{X}}{n}\right)\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{n}\right)\right] \xrightarrow{P} \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$$

  - Por lo tanto, se puede ver como $E(\varepsilon_i^2) = \sigma^2$ (si son independientes), y asumiendo que se tienen momentos finitos (hasta el segundo como mínimo), se puede ver como s^2 converge en probabilidad a σ^2, y eso hace que el estimador $s^2(\boldsymbol{X}'\boldsymbol{X}/n)^{-1}$ sea consistente para $\sigma^2\ (\boldsymbol{X}'\boldsymbol{X}/n)^{-1}$ y que, por tanto, se pueda estimar la varianza asintótica con este

$$s^2 \xrightarrow{P} \sigma^2 \;\Rightarrow\; s^2(\boldsymbol{X}'\boldsymbol{X}/n)^{-1} \xrightarrow{P} \sigma^2\, \boldsymbol{Q}^{-1}$$

$$\Rightarrow Est.Asy.Var\,(\boldsymbol{b}) = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

- Igual que el teorema de Gauss-Markov asegura la eficiencia del estimador de mínimos cuadrados en muestra finita, se pueden definir las condiciones para las cuales el estimador es asintóticamente eficiente

  - Un estimado es asintóticamente eficiente si es consistente, asintóticamente normal y tiene una matriz asintótica de varianzas y covarianzas que no es mayor a la matriz asintótica de varianzas y covarianzas de cualquier otro estimador

- Debido a que se han establecido propiedades de muestra finita y propiedades asintóticas del estimador sin el uso de la suposición de normalidad, entonces se puede decir que el estimado de métodos cuadrados es robusto a violaciones de la suposición de normalidad. En particular, se puede decir que el estimador es robusto a suposiciones de la distribución de $\boldsymbol{\varepsilon}$

  - El estimador de un modelo es robusto si es insensible a desviaciones de las suposiciones básicas del modelo. En términos econométricos, los estimadores robustos mantienen las propiedades deseadas aún sin que se cumplan las suposiciones que motivan el estimado del modelo

    - Por lo tanto, la robustez de un estimador es relativo a una suposición o conjunto de suposiciones concretas, no es un concepto global (sino relativo)

    - La robustez no es necesariamente una característica bien definida para un estimador, por lo que se considera una caracterización amplia de las propiedades asintóticas de ciertas estimaciones y procedimientos

    - En el contexto de las seis suposiciones básicas del modelo de regresión lineal múltiples, se estableció la consistencia del estimador $\boldsymbol{b}$ debido a la suposición de rango completo y a la de independencia de los errores $\boldsymbol{\varepsilon}$ con las $\boldsymbol{X}$. Por lo tanto, otras suposiciones como la normalidad, la heterocedasticidad y la no autocorrelación no son necesarias para obtener las propiedades del estimador de mínimos cuadrados

  - La derivación de la varianza asintótica se basaba en la suposición de la homoscedasticidad y la no autocorrelación. No obstante, se puede hacer una formulación más general para motivar el uso de un estimador robusto a la heterocedasticidad

- Considerando que $Var(\varepsilon_i|\boldsymbol{x_i}) = \sigma_i^2$ (en donde se asume que la variación se debe a $\boldsymbol{x_i}$) se puede obtener la matriz de varianzas y covarianzas del vector de parámetros de mínimos cuadrados ordinarios y su matriz de varianzas y covarianzas asintótica

$$\boldsymbol{b} = \boldsymbol{\beta} + (\boldsymbol{X'X})^{-1} \sum_{\boldsymbol{i}} \boldsymbol{x_i} \varepsilon_i$$

$$\Rightarrow \; Var(\boldsymbol{b}|\boldsymbol{X}) = (\boldsymbol{X'X})^{-1} \left[\sum_{\boldsymbol{i}} \sigma_i^2 \boldsymbol{x_i x_i'} \right] (\boldsymbol{X'X})^{-1}$$

$$\Rightarrow \; Asy.Var(\boldsymbol{b}|\boldsymbol{X}) = \frac{1}{n} \boldsymbol{Q}^{-1} \left[\operatorname{plim}_{n\to\infty} \frac{1}{n} \sum_{\boldsymbol{i}} \sigma_i^2 \boldsymbol{x_i x_i'} \right] \boldsymbol{Q}^{-1} =$$

$$= \frac{1}{n} \boldsymbol{Q}^{-1} \boldsymbol{Q}^{*} \boldsymbol{Q}^{-1}$$

- Una estrategia robusta a la heterocedasticidad no especificada es utilizar un estimador factible de $\boldsymbol{Q}^*$ tal como el estimador robusto a la heterocedasticidad de White. Por lo tanto, se puede estimar la varianza asintótica de manera robusta a la heterocedasticidad no especificada con él

$$\boldsymbol{W_{het}} = \frac{1}{n} \sum_{\boldsymbol{i}} e_i^2 \boldsymbol{x_i x_i'} \quad where \quad e_i = y_i - \boldsymbol{x_i' b}$$

$$Est.Asy.Var(\boldsymbol{b}|\boldsymbol{X}) = n \boldsymbol{Q}^{-1} \boldsymbol{W_{het}} \boldsymbol{Q}^{-1}$$

- De este modo, no se necesita la suposición de homoscedasticidad para obtener los errores estándar apropiados del estimador de mínimos cuadrados ordinarios

- De manera análoga, se puede derivar un estimador de la varianza asintótica cuando en la base de datos hay clústeres o observaciones en grupos

  - Cuando hay grupos de datos muestrales que están relacionados, se supone que hay C grupos o clústeres de observaciones (las cuales se indexan por $c = 1,2, \dots, C$). Las observaciones un clúster se agrupan por las correlaciones entre observaciones de un mismo grupo

  - Asumiendo que hay N_c observaciones en un clúster, entonces $n = \sum_{c=1}^{C} N_c$ y el modelo de regresión es el siguiente:

$$y_{i,c} = \boldsymbol{x_{i,c}'} \boldsymbol{\beta} + \varepsilon_{i,c}$$

- El estimador de mínimos cuadrados con clústeres permite obtener la matriz de varianzas y covarianzas de $\boldsymbol{b}$, donde $\boldsymbol{X_c}$ es la matriz $N_c \times K$ de variables exógenas para el clúster c, $\boldsymbol{\varepsilon_c}$ es el vector de N_c perturbaciones y $\boldsymbol{\Omega_c}$ es una matriz de varianzas y covarianzas del vector completo $\boldsymbol{\varepsilon_c}$ que representa sus correlaciones desestructuradas. Asumiendo que los clústeres son independientes, se obtiene la siguiente fórmula:

$$\boldsymbol{b} = \boldsymbol{\beta} + \left(\sum_{c=1}^{C} \boldsymbol{X_c'X_c}\right)^{-1}\left[\sum_{c=1}^{C}\sum_{i=1}^{N} \boldsymbol{x_{i,c}}\varepsilon_{i,c}\right] = \boldsymbol{\beta} + (\boldsymbol{X'X})^{-1}\left(\sum_{c=1}^{C} \boldsymbol{X_c\varepsilon_c}\right)$$

$$\Rightarrow Var(\boldsymbol{b}|\boldsymbol{X}) = (\boldsymbol{X'X})^{-1}\left(\sum_{c=1}^{C} \boldsymbol{X_c\Omega_c X_c'}\right)(\boldsymbol{X'X})^{-1}$$

- Como antes, se puede encontrar la varianza asintótica y, a partir de esta, encontrar un estimador de la varianza asintótica parecida a la encontrada con el estimador de White

$$Asy.Var(\boldsymbol{b}) = \frac{1}{C}\boldsymbol{Q}^{-1}\left(\operatorname*{plim}_{n\to\infty}\frac{1}{C}\sum_{c=1}^{C} \boldsymbol{X_c\Omega_c X_c'}\right)\boldsymbol{Q}^{-1}$$

$$\boldsymbol{W}_{cluster} = \frac{1}{C}\sum_{c=1}^{C}\left(\sum_{c=1}^{N_c} \boldsymbol{x_{ic}e_c}\right)\left(\sum_{c=1}^{N_c} \boldsymbol{x_{ic}e_c}\right)$$

$$\Rightarrow Est.Asy.Var(\boldsymbol{b}) = C(\boldsymbol{X'X})^{-1}\boldsymbol{W}_{cluster}(\boldsymbol{X'X})^{-1}$$

Los contrastes de hipótesis y la selección del modelo

- Proponiendo un modelo de regresión $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, se puede hacer una pregunta empírica sobre si la especificación del modelo parece ser consistente con los datos a través de los contrastes de hipótesis
    - El enfoque general para contrastar hipótesis es formular un modelo estadístico que contenga las hipótesis como restricciones en sus parámetros
        - Se dice que una teoría económica tiene implicaciones comprobables si implica algunas restricciones contrastables en el modelo. Es decir, si se puede comprobar que uno o más parámetros corresponden a un valor concreto para, así, poder construir una hipótesis nula y contrastarlos

$$\ln I_t = \beta_1 + \beta_2 i_t + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t,$$

$$\ln I_t = \beta_1 + \beta_2 (i_t - \Delta p_t) + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t, \quad \beta_2 + \beta_3 = 0$$

- Por lo tanto, existen modelos que no tienen implicaciones comprobables al no poderse generar una hipótesis nula para los parámetros que se pueda contrastar (ya que una misma variable contiene una combinación de dos o más parámetros, por ejemplo)

$$\ln I_t = \beta_1 + \beta_2 i_t + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t,$$

$$\ln I_t = \beta_1 + \beta_2 (i_t - \Delta p_t) + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \varepsilon_t.$$

  - La descripción de las implicaciones comprobables sugiere que las restricciones contrastables implican que solo algunos de os posibles modelos contenidos en la especificación original serán válidos (consistente con la teoría)

    - El subconjunto de valores que son consistentes con la teoría está contenido dentro de un conjunto de valores no restringido (espacio paramétrico), de modo que los modelos consistentes están anidados y el modelo restringido está contenido en el no restringido. Es decir, a partir del modelo original de la proposición, el otro modelo se puede obtener utilizando restricciones sobre los parámetros de las variables sin eliminar ninguna

    - No obstante, cuando se eliminan variables (restringiendo parámetros a cero), estos modelos no son anidados, aunque se puedan obtener a partir del modelo original (se tienen implicaciones comprobables)

- Para poder contrastar hipótesis, se consideran tres enfoques diferentes: los contrastes de Wald, los contrastes basados en el ajuste y los contrastes de multiplicadores de Lagrange

  - La lógica que sigue cada tipo de procedimiento o contraste es la siguiente:

    - En los contrastes de Wald, la hipótesis expresa que $\boldsymbol{\beta}$ obedece algunas restricciones, las cuales se pueden expresar como $\boldsymbol{c}(\boldsymbol{\beta}) = \mathbf{0}$. Siendo $\boldsymbol{b}$ el estimador de mínimos cuadrados consistente de $\boldsymbol{\beta}$, el contraste de Wald mide que tan cerca $\boldsymbol{c}(\boldsymbol{\beta})$ está de cero, basándose en la estimación del modelo no restringido (que tan cerca está el modelo sin restricciones de las restricciones hipotéticas)

- El mejor ajuste posible (mayor R^2) se obtiene usando mínimos cuadrados sin poner restricciones, pero añadir restricciones degrada el ajuste del modelo. En los contrastes basados en el ajuste, se mide cuando decrece R^2 cuando se imponen restricciones, comparando el ajuste del modelo restringido con el no restringido

- El contraste de multiplicadores de Lagrange se basa en el modelo restringido, y se basa en un resultado general sobre que, cuando se imponen restricciones, si estas son incorrectas, es posible detectar el fallo con estadísticos medibles. Por lo tanto, se basa en estimadores del modelo restringido (por ejemplo, correlación entre residuos y otros)

- Para desarrollar los procedimientos de contrastes de hipótesis, se asume homocedasticidad y perturbaciones normalmente distribuidas, dado que permiten saber las distribuciones exactas de los estadísticos de contraste

  - No obstante, más adelante se desarrollan resultados alternativos para poder proceder sin las dos suposiciones

  - Es importante distinguir entre la teoría subyacente de los procedimientos de contraste de las mecánicas prácticas de inferencia, basadas en aproximaciones asintóticas y matrices de varianzas y covarianzas robustas

- La hipótesis lineal general de un conjunto de J restricciones de un modelo de regresión lineal $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ se puede escribir de la siguiente manera:

$$\begin{aligned} r_{11}\beta_1 + r_{12}\beta_2 + \cdots + r_{1K}\beta_K &= q_1 \\ r_{21}\beta_1 + r_{22}\beta_2 + \cdots + r_{2K}\beta_K &= q_2 \\ &\cdots \\ r_{J1}\beta_1 + r_{J2}\beta_2 + \cdots + r_{JK}\beta_K &= q_J. \end{aligned}$$

  - El caso general se puede escribir en notación matricial como $\boldsymbol{R\beta} = \boldsymbol{q}$, donde cada fila de $\boldsymbol{R}$ son los coeficientes de una de las restricciones. Típicamente, $\boldsymbol{R}$ solo tendrá unas pocas filas y muchos ceros en cada fila

A set of the coefficients sum to one, $\beta_2 + \beta_3 + \beta_4 = 1$,

$$\mathbf{R} = [0 \quad 1 \quad 1 \quad 1 \quad 0 \quad \cdots]; \mathbf{q} = 1.$$

A subset of the coefficients are all zero, $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_3 = 0$,

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} = [\mathbf{I} \mid \mathbf{0}]; \quad \mathbf{q} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Several linear restrictions, $\beta_2 + \beta_3 = 1$, $\beta_4 + \beta_6 = 0$, and $\beta_5 + \beta_6 = 0$,

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}; \quad \mathbf{q} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

All the coefficients in the model except the constant term are zero,

$$\mathbf{R} = [\mathbf{0} | \mathbf{I}_{K-1}]; \quad \mathbf{q} = \mathbf{0}.$$

- La matriz $\boldsymbol{R}$ tiene K columnas (para coincidir con las filas de $\boldsymbol{\beta}$), J filas para un total de J restricciones, y tiene rango completo de filas, de modo que J tiene que ser menor o igual a K

- No obstante, $K \neq J$, ya que, de no ser así, la matriz sería cuadrada e invertible y $\boldsymbol{\beta} = \boldsymbol{R^{-1}q}$ (no habría problema de estimación o inferencia porque se ha dado un valor para cada parámetro), haciendo que $J < K$. La restricción $\boldsymbol{R\beta} = \boldsymbol{q}$ impone J restricciones en los K parámetros, haciendo que hayan $K - J$ parámetros libres

  - Para poder extender estos métodos a restricciones no lineales, se plantea una hipótesis no lineal general, que involucra un conjunto $\boldsymbol{c(\beta)}$ de J restricciones no lineales posibles de $\boldsymbol{\beta}$

$$\boldsymbol{c(\beta) = q}$$

- La contraparte de los requerimientos vistos para el caso especial en donde $\boldsymbol{c(\beta)}$ es no lineal es que $J < K$ y que la matriz jacobiana de estas restricciones (formada por el gradiente de $\boldsymbol{c(\beta)}$ con respecto a cada uno de los parámetros en $\boldsymbol{\beta}$) tenga rango de fila completo

- En el caso lineal, la jacobiana equivale a $\boldsymbol{R}$ y la independencia funcional que se requiere sería equivalente a la independencia lineal

- El contraste de Wald es el procedimiento más utilizado, y normalmente se llama contraste de significación. El principio operador de este procedimiento es ajustar la regresión sin las restricciones y, entonces, evaluar si los resultados parecen, dentro de la variabilidad muestral, ser consistentes con la hipótesis

- El caso más simple es el de contrastar un valor para un solo coeficiente. La distancia de Wald del coeficiente estimado respecto a su valor hipotético es la distancia medida en unidades de desviación estándar, por lo que se puede expresar a través del siguiente estadístico:

$$\begin{cases} H_0: \beta_k = \beta_k^0 \\ H_1: \beta_k \neq \beta_k^0 \end{cases} \quad \Rightarrow \quad W_k = \frac{b_k - \beta_k^0}{\sqrt{\sigma^2 S^{kk}}}$$

$$where \;\; S^{kk} = k^{th}\; diagonal\; elem.\; of\; (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

  - El estadístico W_k sigue una distribución normal estándar asumiendo que $E(b_k) = \beta_k^0$. Si no es igual a β_k^0, el estadístico sigue teniendo una distribución normal, pero la media no será nula. Si, en particular, $E(b_k) = \beta_k^1$ (diferente de β_k^0), entonces se obtiene la siguiente igualdad:

$$E[W_k | E(b_k) = \beta_k^1] = \frac{E(b_k) - \beta_k^0}{\sqrt{\sigma^2 S^{kk}}} = \frac{\beta_k^1 - \beta_k^0}{\sqrt{\sigma^2 S^{kk}}}$$

  - Con tal de utilizar W_k para contrastes de hipótesis, la interpretación es que si β_k es igual a β_k^0, entonces b_k estará cerca de β_k^0 (con la distancia medida en unidades de errores estándar). Por lo tanto, la lógica del contraste es rechazar H_0 si W_k es grande en valor absoluto

- Sin embargo, no se puede usar la medida de Wald propuesta porque σ^2 no se sabe, de modo que se utiliza el estimador s^2 para así obtener un estadístico t

$$\begin{cases} H_0: \beta_k = \beta_k^0 \\ H_1: \beta_k \neq \beta_k^0 \end{cases} \quad \Rightarrow \quad t_k = \frac{b_k - \beta_k^0}{\sqrt{s^2 S^{kk}}}$$

$$where \;\; S^{kk} = k^{th}\; diagonal\; elem.\; of\; (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

  - Asumiendo que β_k es igual a β_k^0, el estadístico sigue una distribución t-Student con $n - K$ grados de libertad

  - A partir de este estadístico, se puede construir un procedimiento de contraste de hipótesis a través de determinar un nivel de confianza (para el intervalo de confianza) y de significación (para el contraste)

$$P\left(t_{n-K,\frac{\alpha}{2}} < \; t_k < t_{n-K,1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$P\left(t_k < t_{n-K,1-\alpha}\right) = 1 - \alpha \quad or \quad P\left(t_k > t_{n-K,\alpha}\right) = 1 - \alpha$$

- Considerando que se quiere contrastar un conjunto de J restricciones lineales en la hipótesis nula contras las de la hipótesis alternativa, se puede utiliza el criterio de Wald

$$\begin{cases} H_0: \boldsymbol{R\beta} - \boldsymbol{q} = \boldsymbol{0} \\ H_1: \boldsymbol{R\beta} - \boldsymbol{q} \neq \boldsymbol{0} \end{cases}$$

  - Dado el estimador de mínimos cuadrados $\boldsymbol{b}$, el interés se centra en el vector de discrepancia $\boldsymbol{Rb} - \boldsymbol{q} = \boldsymbol{m}$. Es improbable que $\boldsymbol{m} = \boldsymbol{0}$, pero es razonable preguntarse si la desviación se atribuye a la variabilidad muestral o si es significativa

  - Como $\boldsymbol{b}$ sigue una distribución normal y $\boldsymbol{m}$ es una función lineal de $\boldsymbol{b}$, $\boldsymbol{m}$ también se distribuye normalmente. Por lo tanto, si la hipótesis nula es verdad, entonces se obtienen los siguientes resultados:

$$E(\boldsymbol{m}|\boldsymbol{X}) = \boldsymbol{R}E(\boldsymbol{b}|\boldsymbol{X}) - \boldsymbol{q} = \boldsymbol{R\beta} - \boldsymbol{q} = \boldsymbol{0}$$

$$Var(\boldsymbol{m}|\boldsymbol{X}) = \boldsymbol{R}Var(\boldsymbol{b}|\boldsymbol{X})\boldsymbol{R}' = \boldsymbol{R}[\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]\boldsymbol{R}'$$

$$\Rightarrow \ \boldsymbol{Rb} \sim N(\boldsymbol{q}, \boldsymbol{R}[\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]\boldsymbol{R}')$$

- En consecuencia, es posible basarse en el criterio de Wald para contrastar H_0. Condicionando en $\boldsymbol{X}$, se puede comprobar que el estadístico W sigue una distribución χ_J^2 si la hipótesis nula es cierta, y así ver como valores grandes del estadístico indican que la hipótesis nula es rechazable (la distancia entre $\boldsymbol{b}$ y $\boldsymbol{\beta}$ es muy grande)

$$W = \boldsymbol{m}'Var(\boldsymbol{m}|\boldsymbol{X})^{-1}\boldsymbol{m} =$$

$$= (\boldsymbol{Rb} - \boldsymbol{q})'[\boldsymbol{R}[\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]\boldsymbol{R}']^{-1}(\boldsymbol{Rb} - \boldsymbol{q}) \sim \chi_J^2$$

  - Este estadístico, no obstante, no se puede usar porque depende de σ^2, por lo que se puede obtener un estadístico F haciendo esa sustitución y dividendo por los grados de libertad

$$F = \frac{W}{J}\frac{\sigma^2}{s^2} = \frac{(\boldsymbol{Rb} - \boldsymbol{q})'[\boldsymbol{R}[s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]\boldsymbol{R}']^{-1}(\boldsymbol{Rb} - \boldsymbol{q})}{J} \sim F_{J,n-K}$$

  - Para comprobar una sola restricción lineal de la forma $\boldsymbol{r}'\boldsymbol{\beta} = q$, el estadístico F será el siguiente:

$$H_0: r_1\beta_1 + r_2\beta_2 + \cdots + r_K\beta_K = \mathbf{r}'\boldsymbol{\beta} = q$$

$$F[1, n-K] = \frac{\left(\sum_j r_j b_j - q\right)^2}{\sum_j \sum_k r_j r_k Est.Var.\left(b_j, b_k\right)}$$

- Para contrastar la hipótesis de que el coeficiente j es igual a un valor particular, entonces $\boldsymbol{R}$ tiene una sola fila con un 1 en la posición j, $\boldsymbol{R}[s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]\boldsymbol{R}'$ es el elemento diagonal j de la matriz de varianzas y covarianzas estimada y $\boldsymbol{Rb} - \boldsymbol{q}$ es $(b_j - q)$. El estadístico, por tanto, será el siguiente:

$$F[1, n-K] = \frac{(b_j - q)^2}{Est.Var.(b_j)}$$

- Considerando un enfoque alternativo, se puede plantear una estimación muestral $\boldsymbol{r}'\boldsymbol{b} = \hat{q}$ para $\boldsymbol{r}'\boldsymbol{\beta} = q$. Si $\hat{q}$ difiere significativamente de q, se puede concluir que los datos muestrales no son consistentes con la hipótesis nula, de modo que se puede usar el estadístico t

$$t = \frac{\hat{q} - q}{\sqrt{Est.Var(\hat{q}|\boldsymbol{X})}}$$

$$where\ \ Est.Var(\hat{q}|\boldsymbol{X}) = \boldsymbol{r}'[s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]\boldsymbol{r}$$

- Existe una relación útil entre el estadístico t y el F, ya que es posible escribir el cuadrado del estadístico t en términos del estadístico F cuando hay una sola restricción

$$t^2 = \frac{(\hat{q} - q)^2}{Est.Var(\hat{q} - q|\boldsymbol{X})} =$$

$$= \frac{(\boldsymbol{r}'\boldsymbol{b} - q)'[\boldsymbol{r}[s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]\boldsymbol{r}']^{-1}(\boldsymbol{r}'\boldsymbol{b} - q)}{1} \sim F_{1,n-K}$$

- Un enfoque diferente se enfoca en utilizar el ajuste de la regresión, de modo que uno se puede preguntar si escoger valores alternativos para los parámetros permite obtener una pérdida significativa de ajuste
    - El vector de coeficientes $\boldsymbol{b}$ es escogido para minimizar la suma de desviaciones al cuadrado $\boldsymbol{e}'\boldsymbol{e}$
        - Debido a que R^2 iguala $1 - \boldsymbol{e}'\boldsymbol{e}/\boldsymbol{y}'\boldsymbol{M^0}\boldsymbol{y}$ y $\boldsymbol{y}'\boldsymbol{M^0}\boldsymbol{y}$ es una constante que no involucra a $\boldsymbol{b}$, si el modelo contiene un término constante, $\boldsymbol{b}$ se escoge para maximizar R^2
        - En consecuencia, se puede construir un estadístico de contraste que se base en comparar el R^2 entre dos regresiones

- Suponiendo que se impone explícitamente las restricciones de la hipótesis lineal general en la regresión, es posible obtener un estimador restringido $\boldsymbol{b}_*$ a través de utilizar la función lagrangiana

$$\min_{\boldsymbol{b_0}} \boldsymbol{e'e} = (\boldsymbol{y} - \boldsymbol{Xb_0})'(\boldsymbol{y} - \boldsymbol{Xb_0}) \;\; s.t. \;\; \boldsymbol{Rb_0} = \boldsymbol{q}$$

$$\min_{\boldsymbol{b_0}} L^*(\boldsymbol{b_0}, \boldsymbol{\lambda}) = (\boldsymbol{y} - \boldsymbol{Xb_0})'(\boldsymbol{y} - \boldsymbol{Xb_0}) + 2\boldsymbol{\lambda}'(\boldsymbol{Rb_0} - \boldsymbol{q})$$

$$\frac{\partial L^*}{\partial \boldsymbol{b}_*} = -2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{Xb}_*) + 2\boldsymbol{R}'\boldsymbol{\lambda}_* = \boldsymbol{0} \qquad \frac{\partial L^*}{\partial \boldsymbol{\lambda}_*} = 2(\boldsymbol{Rb_0} - \boldsymbol{q}) = \boldsymbol{0}$$

  - Dividiendo entre 2 (por eso se ha multiplicado $\boldsymbol{\lambda}$ por 2) y expandiendo los términos, se puede obtener la siguiente ecuación matricial:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{R}' \\ \mathbf{R} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b}_* \\ \boldsymbol{\lambda}_* \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{q} \end{bmatrix}$$

  - Asumiendo que la matriz partida es invertible, se tiene la siguiente solución:

$$\begin{bmatrix} \mathbf{b}_* \\ \boldsymbol{\lambda}_* \end{bmatrix} = \begin{bmatrix} \mathbf{X'X} & \mathbf{R}' \\ \mathbf{R} & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X'y} \\ \mathbf{q} \end{bmatrix} = \mathbf{A}^{-1}\mathbf{d}.$$

  - Si además $\boldsymbol{X'X}$ es invertible, entonces se tiene soluciones explícitas:

$$\boldsymbol{b}_* = \boldsymbol{b} - (\boldsymbol{X'X})^{-1}\boldsymbol{R}'[\boldsymbol{R}(\boldsymbol{X'X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{Rb} - \boldsymbol{q}) = \boldsymbol{b} - \boldsymbol{Cm}$$

$$where \;\; \boldsymbol{C} = (\boldsymbol{X'X})^{-1}\boldsymbol{R}'[\boldsymbol{R}(\boldsymbol{X'X})^{-1}\boldsymbol{R}']^{-1}$$

$$\boldsymbol{\lambda}_* = [\boldsymbol{R}(\boldsymbol{X'X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{Rb} - \boldsymbol{q})$$

  - Como $\boldsymbol{\lambda}_*$ involucra el vector de discrepancia $\boldsymbol{m} = \boldsymbol{Rb} - \boldsymbol{q}$, si se satisface la restricción, entonces $\boldsymbol{b}_* = \boldsymbol{b}$ y el estimador no tendría ningún sesgo si las hipótesis son ciertas

- La solución obtenida por mínimos cuadrados restringidos cumple unas propiedades interesantes:

  - El estimador obtenido es lineal, por lo que se cumple $\boldsymbol{Rb}_* - \boldsymbol{q}$

  - Bajo normalidad, el estimador obtenido es igual al estimador de máxima verosimilitud $\boldsymbol{b_{MLE}} = \boldsymbol{b}_*$. Por lo tanto, este sigue una distribución normal multivariante y es consistente y asintóticamente eficiente

- La matriz de varianzas y covarianzas para $\boldsymbol{b}_*$ es simplemente σ^2 veces el bloque superior de la matriz partida $\boldsymbol{A}^{-1}$. Si $\boldsymbol{X}'\boldsymbol{X}$ es invertible, se puede hacer una formulación explícita:

$$Var(\boldsymbol{b}_*|\boldsymbol{X}) = Var(\boldsymbol{b}|\boldsymbol{X}) - nonnegative\ def.matrix =$$

$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} - \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

- Una interpretación para esta reducción en la varianza es como el valor de la información contenida en las restricciones. La matriz $Var(\boldsymbol{b}_*|\boldsymbol{X})$ es más pequeña que $Var(\boldsymbol{b}|\boldsymbol{X})$ aunque las restricciones sean incorrectas (dado que añadir parámetros hace que aumente la variabilidad de las estimaciones), por lo que el estimador de la varianza es la varianza mínima entre los estimadores lineales y sin sesgo que cumplen el conjunto de restricciones

- Para desarrollar un contraste basado en el estimador de mínimos cuadrados restringido, se puede utilizar el resultado anterior sobre la relación entre el cuadrado del estadístico t y el estadístico F para J restricciones

  - El ajuste de los coeficientes de mínimos cuadrados restringidos no puede ser mejores que la solución no restringida (debido a que se considera un espacio paramétrico más pequeño), lo cual se puede demostrar a través de la suma cuadrada de desviaciones:

$$\boldsymbol{e}_* = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} - \boldsymbol{X}(\boldsymbol{b}_* - \boldsymbol{b}) = \boldsymbol{e} - \boldsymbol{X}(\boldsymbol{b}_* - \boldsymbol{b})$$

$$\Rightarrow \boldsymbol{e}_*'\boldsymbol{e}_* = (\boldsymbol{e}' - (\boldsymbol{b}_* - \boldsymbol{b})'\boldsymbol{X}')\big(\boldsymbol{e} - \boldsymbol{X}(\boldsymbol{b}_* - \boldsymbol{b})\big) =$$

$$= \boldsymbol{e}'\boldsymbol{e} - \boldsymbol{e}'\boldsymbol{X}(\boldsymbol{b}_* - \boldsymbol{b}) - (\boldsymbol{b}_* - \boldsymbol{b})'\boldsymbol{X}'\boldsymbol{e} + (\boldsymbol{b}_* - \boldsymbol{b})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b}_* - \boldsymbol{b}) =$$

$$= \boldsymbol{e}'\boldsymbol{e} + (\boldsymbol{b}_* - \boldsymbol{b})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b}_* - \boldsymbol{b}) \geq \boldsymbol{e}'\boldsymbol{e}$$

$$\Rightarrow \boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{b}_* - \boldsymbol{b})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b}_* - \boldsymbol{b}) =$$

$$= (\boldsymbol{R}\boldsymbol{b} - \boldsymbol{q})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\boldsymbol{b} - \boldsymbol{q})$$

  - La expresión obtenida para la pérdida de ajuste $\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e}$ aparece en el numerador el estadístico F, por lo que se puede sustituir y así encontrar una expresión equivalente

$$\frac{(\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e})/J}{\boldsymbol{e}'\boldsymbol{e}/(n-K)} \sim F_{J,n-K}$$

- Finalmente, dividiendo el numerador y el denominador por la desviación cuadrada de la media muestral $\sum_i (y_i - \bar{y})^2$ se puede obtener el estadístico en términos de R^2

$$\frac{(R^2 - R_*^2)/J}{(1 - R^2)/(n - K)} \sim F_{J,n-K}$$

- Una de las preguntas más importantes es normalmente ver si la ecuación de la regresión es significativa en su totalidad. Este es un contraste conjunto de las hipótesis de que todos los coeficientes (excepto el término constante) sean nulos

  - Si todos los coeficientes son nulos, entonces el coeficiente de determinación es nulo también, de modo que se puede basar el contraste en la hipótesis sobre el valor de $R_*^2 = 0$. Por lo tanto, el estadístico será el siguiente:

$$\frac{R^2/(K - 1)}{(1 - R^2)/(n - K)} \sim F_{J,n-K}$$

  - Valores grandes de F evidencian que la hipótesis nula se puede rechazar, y estos son inducidos por valores grandes de R^2, por lo que la lógica del contraste es que F mide la pérdida de ajuste que resulta en imponer la restricción de que $\boldsymbol{\beta} = \mathbf{0}$

Las variables binarias o categóricas y los experimentos

- Una de las herramientas más útiles en el análisis de regresión son las variables binarias, las cuales son convenientes para crear desplazamientos discretos de la función en un modelo de regresión

  - Las variables binarias toman el valor de 1 para algunas observaciones para indicar la presencia de un efecto o la membresía a un grupo y toma el valor 0 para el resto de observaciones

    - Las variables binarias normalmente se usan en ecuaciones de regresión que contienen otras variables cuantitativas

    - Una variable binaria que toma el valor de 1 solo para una única observación tiene el efecto de eliminar esa observación del cálculo de los parámetros de mínimos cuadrados y el estimador de la varianza (pero no del R^2)

  - El grupo de observaciones a las que se les atribuye el 0 se denomina grupo de referencia o *benchmark group*, y la interpretación del coeficiente se hace relativa a este grupo (como una diferencia)

- Por lo tanto, en un modelo con una sola variable binaria, el cambio producido por la variable binaria d se debe interpretar como el efecto de pertenecer a un grupo o que ocurra un evento (las observaciones para las cuales $d = 1$) frente a no pertenecer o a que no ocurra (grupo de referencia)

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \varepsilon_i$$

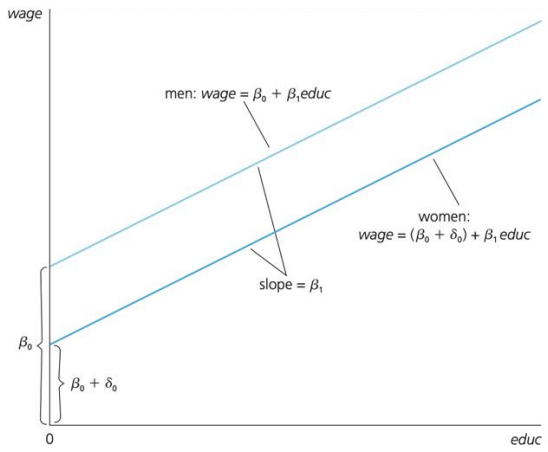
$$\Rightarrow \frac{\partial E(y_i|x_i, d_i)}{\partial d_i} = \beta_2 = E(y_i|x_i, d_i = 1) - E(y_i|x_i, d_i = 0)$$

- Como se puede deducir matemáticamente, el efecto de un cambio en d solo depende de d y no de cualquier otra variable (al mantenerse constante)

- En un modelo de este tipo, esto se traduce en un desplazamiento del intercepto y, por tanto, de la función de regresión

$$E(y_i|x_i, d_i = 1) = (\beta_0 + \beta_2) + \beta_1 x_i + \varepsilon_i$$

$$E(y_i|x_i, d_i = 0) = \beta_0 + \beta_1 x_i + \varepsilon_i$$



- Debido al teorema de Frisch-Waugh-Lovell, la regresión múltiple sin coeficiente constante es equivalente a hacer una regresión de la variable dependiente y de las variables independientes sobre las variables binarias y hacer una regresión de los residuos de la primera sobre los residuos de la segunda

  - Por lo tanto, se puede ver que el vector de coeficientes de las variables cuantitativas seguirá siendo el mismo y se produce el mismo vector de residuos que cuando se incluye el conjunto de variables binarias en la regresión. Esto ocurre porque, si hay $K - r$ variables continuas y r variables binarias, entonces la matriz X

tiene un tamaño $n \times K$ y, debido a que se añaden todas las variables binarias (para que no haya problemas de identificación), entonces hay columnas de ceros y unos que sumadas permiten obtener una columna de unos, y por tanto es como tener un parámetro constante

- Los valores de la variable binaria son arbitrarios, dado que se puede trabajar con cualquier otro par de números al ser combinaciones lineales de 1 y 0, de modo que la codificación no es relevante para la interpretación

$$d = \begin{cases} 1 \ \ if \ event \ A \\ 0 \ \ if \ event \ B \end{cases} \quad g = \begin{cases} a \ \ if \ event \ A \\ b \ \ if \ event \ B \end{cases} \ \Rightarrow \ g = ad + b$$

  - Cuando cambia la codificación de la variable binaria también cambia el coeficiente (debido a la magnitud diferente de los datos). No obstante, como $\boldsymbol{y}$ no cambia, entonces se puede plantear la siguiente ecuación para cambiar la codificación

$$E(y_i|x_i, d_i) = \beta_0 + \beta_1 x_i + \beta_2 d_i = \beta_0^* + \beta_1 x_i + \beta_2^* g_i = E(y|x, g)$$

$$E(y_i|x_i, d_i = 1) - E(y_i|x_i, d_i = 0) = E(y|x, d = a) - E(y|x, d = b)$$

  - Obviamente, cuando se cambia el grupo base, la codificación es la inversa y los coeficientes serán exactamente los contrarios

$$\beta_2 = E(y_i|x_i, d_i = 1) - E(y_i|x_i, d_i = 0)$$

$$\beta_2^* = E(y_i|x_i, d_i = 0) - E(y_i|x_i, d_i = 1)$$

$$\Rightarrow \ \beta_2 = \beta_2^* = -\beta_2$$

  - Finalmente, como $\boldsymbol{y}$ no cambia, aunque se cambie la codificación, los modelos de regresión tienen que dar el mismo resultado, por lo que se puede ver que el coeficiente constante cambiará cuando la codificación cambie

$$E(y_i|x_i, d_i) = \beta_0 + \beta_1 x_i + \beta_2 d_i = \beta_0^* + \beta_1 x_i + \beta_2^* g_i = E(y_i|x_i, g_i)$$

$$E(y_i|x_i, d_i = 1) \Rightarrow \ \beta_0^* = \beta_0 - \beta_2 + \beta_2^* a$$

$$E(y_i|x_i, d_i = 0) \ \ \Rightarrow \ \ \beta_0^* = \beta_0 + \beta_2^* b$$

$$\beta_2^* = -\frac{\beta_2}{b - a}$$

- En el caso en el que se incluya más de una variable binaria, la interpretación de los coeficientes sigue siendo la misma (la diferencia con el grupo de referencia)

y el efecto gráfico de estas también. No obstante, hay algunas diferencias con el caso anterior

- En la mayoría de aplicaciones, las variables binarias se utilizan para factores puramente cualitativos, pero hay casos en los que se utilizan para representar niveles de un factor subyacente que podría haber sido medido si fuera posible (en donde hay umbrales o categorías)

  - En este caso, se pueden dividir los diferentes umbrales o categorías a través de varias variables binarias

$$income = \beta_1 + \beta_2 age + \delta_B B + \delta_M M + \delta_P P + \varepsilon$$

  - Sin embargo, la definición del grupo de referencia tendrá un efecto en la función de regresión, dado que se puede hacer que 1 sea pertenecer a una categoría concreta o que sea la categoría más alta (o baja, depende del criterio)

$$\begin{aligned}
&\text{High school:} && E[income \mid age, HS] = \beta_1 + \beta_2\, age,\\
&\text{Bachelor's:} && E[income \mid age, B] = \beta_1 + \beta_2\, age + \delta_B,\\
&\text{Master's:} && E[income \mid age, M] = \beta_1 + \beta_2\, age + \delta_M,\\
&\text{Ph.D.:} && E[income \mid age, P] = \beta_1 + \beta_2\, age + \delta_P.
\end{aligned}$$

$$\begin{aligned}
&\text{High school:} && E[income \mid age, HS] = \beta_1 + \beta_2\, age,\\
&\text{Bachelor's:} && E[income \mid age, B] = \beta_1 + \beta_2\, age + \delta_B,\\
&\text{Master's:} && E[income \mid age, M] = \beta_1 + \beta_2\, age + \delta_B + \delta_M,\\
&\text{Ph.D.:} && E[income \mid age, P] = \beta_1 + \beta_2\, age + \delta_B + \delta_M + \delta_P.
\end{aligned}$$

  - Eso, a su vez, cambia la interpretación de estos coeficientes, dado que, en el primer caso, el coeficiente representa la diferencia entre pertenecer a la categoría y pertenecer al grupo base, y en el segundo caso, representa el efecto marginal de pertenecer a una categoría (dado que también se puede pertenecer a otras)

  - Como se puede observar, las observaciones pertenecen al grupo de referencia cuando todas las variables binarias incluidas son nulas

- Cuando hay varias categorías dentro de las observaciones, el uso de variables binarias es necesario

  - No obstante, el modelo de regresión no se incluyen variables binarias para todas las categorías debido a que, de ser así, entonces la suma de estas sería igual a 1 y eso haría que se

reprodujera el término constante (las columnas de las variables binarias son combinación lineal de la columna de 1 del coeficiente constante), de modo que no habría rango máximo (habría multicolinearidad perfecta)

- A este hecho se le denomina trampa de las variables binarias, y para poder evitar este problema, es necesario no incluir una de las variables binarias o desestacionalizar los datos a través de quitar el término constante

- El caso en que varios grupos de variables binarias se necesita es casi el mismo que cuando solo se necesita un grupo, pero con una importante excepción

  - Como antes, es necesario quitar una de las variables binarias en el grupo para que este no sea combinación lineal del regresor constante del término constate

  - No obstante, cuando hay r grupos de k variables binarias en una regresión, la suma de las variables binarias para cada uno de los r grupos es 1, por lo que hay que omitir una de las variables en cada uno de los r grupos (de modo que se incluya $k-1$ en cada grupo)

- Como los valores de las variables binarias son arbitrarios, la codificación tampoco es relevante para la interpretación en este caso

  - Cuando cambia la codificación de las variables binarias también cambia el coeficiente (debido a la magnitud diferente de los datos). No obstante, como $\boldsymbol{y}$ no cambia, entonces se puede plantear la ecuación anteriormente vista para cambiar la codificación

  - No obstante, cuando en un grupo de variables binarias se cambia el grupo de referencia, entonces no solo se modifica uno de los coeficientes, sino todos los de las variables binarias del mismo grupo. Por tanto, como cada coeficiente era la diferencia con respecto al grupo de referencia anterior, se pueden utilizar esas diferencias para poder ver que los nuevos coeficientes serán las diferencias de los coeficientes originales

  - Obviamente, el coeficiente constante también cambiará y se puede plantear una ecuación como la anteriormente vista para poder obtener el nuevo coeficiente constante

La no linealidad en las variables

- Las variables binarias son particularmente útiles en el modelo de regresión log-lineal para poder saber cual es el cambio porcentual al pasar del grupo de referencia a otro

$$100\%\left(\frac{\partial E(y_i|x_i\,,d_i)}{\partial d}\right)=100\%(\beta_2)=100\%\left(e^{\beta_2}-1\right)$$

  - La segunda expresión equivalente para el efecto marginal del cambio en d se puede obtener a través de la discretización del cambio y de expresar la ecuación en términos de exponenciales

$$100\%\left(\frac{\Delta E(y_i|x_i,d_i)}{\Delta d_i}\right)=100\%\left(\frac{E(y_i|x_i,d_i=1)-E(y_i|x_i,d_i=0)}{E(y_i|x_i,d_i=0)}\right)=$$

$$=100\%\left(\frac{e^{\beta_0+\beta_1x_i+\beta_2}E(e^{\varepsilon})-e^{\beta_0+\beta_1x_i}E(e^{\varepsilon})}{e^{\beta_0+\beta_1x_i}E(e^{\varepsilon})}\right)=100\%\left(e^{\beta_2}-1\right)$$

La evaluación de los efectos de tratamientos

- La principal aplicación de los modelos de selección y de endogeneidad es la evaluación de efectos de tratamientos, que se centra en analizar el efecto de la participación en un tratamiento T en una variable de resultado y. El objetivo principal de analizar el tratamiento o la intervención es saber el efecto del tratamiento en los tratados, pero se presentan varios problemas

  - El analista tiene el riesgo de atribuir al tratamiento efectos causales que deberían ser atribuidos a factores que motivan el tratamiento y el resultado a la vez (endogeneidad del tratamiento)

    - Los ejemplos más comunes suelen ser la raza, la edad y otras características de los individuos

  - Normalmente falta un contrafactual, dado que hay efectos que se quieren medir pero que, por la naturaleza de los individuos, no se puede porque solo se puede observar un estado

    - Un ejemplo claro son los experimentos sobre el efecto de la educación, en los que se querría medir el efecto en la vida de un individuo con educación y sin ella, pero solo se puede observar uno de los dos estados

- El modelo causal de Rubin proporciona un marco útil para el análisis. En este, cada individuo de una población tiene un resultado potencial y y puede estar expuesto a un tratamiento C (C_i es el indicador de si se ha recibido el tratamiento o no). Por lo tanto, hay dos resultados potenciales: $y_i|(C_i = 1) = y_{i1}$ y $y_i|(C_i = 0) = y_{i0}$

- El efecto medio del tratamiento o *average treatment effect* (ATE) es la diferencia media entre los resultados potenciales para toda la población. No obstante, el individuo existe en uno de los dos estados, por lo que no es posible estimar el ATE y se pone más interés en el ATE de los tratados (ATET)

$$ATE = E(y_{i1} - y_{i0}) \qquad ATET = E(y_{i1} - y_{i0}|C_i = 1)$$

- El reto de la investigación actualmente está en encontrar estimadores robustos que no se apoyen mucho en suposiciones frágiles como la identificación por la forma funcional o por exclusión de restricciones. Por ello, $\boldsymbol{x}$ será la información exógena con la que se tiene que lidiar en el problema de estimación, y se hacen suposiciones sobre este vector de variables como las siguientes:

  - Recibir el tratamiento C_i no depende del resultado una vez que los efectos de $\boldsymbol{x}$ se tienen en cuenta. Por lo que, si la asignación de individuos al grupo de tratamiento es aleatoria, entonces se tendría que omitir el efecto de $\boldsymbol{x}$ en la suposición de la independencia condicional

  - La suposición de independencia condicional se extiende a enfoques de regresión con la suposición de media condicional o *conditional mean assumption*, de modo que los resultados de los individuos que no reciben el tratamiento no afectan a la participación

$$E(y_{i0}|\boldsymbol{x_i}, C_i = 1) = E(y_{i0}|\boldsymbol{x_i}, C_i = 0) = E(y_{i0}|\boldsymbol{x_i})$$

  - La suposición sobre la distribución de resultados potenciales expresa que el modelo que se usa para los resultados potenciales de los tratados y los no tratados es el mismo. En el contexto de una regresión eso quiere decir que la misma regresión aplica para ambos grupos y que las perturbaciones no tienen correlación con T (T es exógena). Esta suposición es fuerte, pero permite evitar la endogeneidad

$$f(y|\boldsymbol{x}, T = 1) = f(y|\boldsymbol{x}, T = 0)$$

  - La suposición de superposición expresa que paara cualquier valor $\boldsymbol{x}$, $0 < P(C_i = 1|\boldsymbol{x}) < 1$, de modo que las desigualdades

estrictas expresan que la población contiene una mezcla de individuos tratados y no tratados para cualquier $\boldsymbol{x}$. Esta suposición permite suponer que, de media, para cada individuo tratado habrá otro individuo no tratado con las mismas características (útil para la regresión)

- Para poder analizar el efecto de un tratamiento sobre una respuesta, se puede formular un modelo de regresión con una variable binaria C con valor 1 si se recibe el tratamiento y 0 si no y otras variables $\boldsymbol{x}$

$$y_i = \boldsymbol{x_i'\beta} + \delta C_i + \varepsilon_i$$

  - En este caso, δ es el parámetro de desplazamiento de la función de regresión y mide el impacto del tratamiento en la respuesta de los individuos de la muestra (condicionado a las $\boldsymbol{x}$)

- Cuando el tratamiento o el análisis de este ocurre a lo largo del tiempo, entonces la variable binaria será $C_t = 0$ en el primer periodo y $C_t = 1$ en el segundo, y el efecto δ medirá el cambio en la variable de respuesta antes y después del tratamiento

  - Como la suposición de que $\boldsymbol{x}$ se mantiene constante entre periodos debilita la comparación, una estrategia es incluir un grupo de observaciones de control que no reciban tratamiento, de modo que el cambio en el grupo de tratamiento se puede comparar al cambio del grupo no tratado bajo la suposición de que la diferencia se debe al tratamiento

  - El impacto en este caso se mide con la diferencia de las diferencias anteriormente descritas, de modo que el estimado del parámetro de C será el estimador de diferencias en diferencias

$$\delta = E(\Delta y_i | \Delta C_i = 1) - E(\Delta y_i | C_i = 0) =$$

$$= [E(y_{i1} | C_i = 1) - E(y_{i0} | C_i = 1)] - [E(y_{i1} | C_i = 0) - E(y_{i0} | C_i = 0)]$$

- Cuando el tratamiento es el resultado de un evento que ocurre completamente fuera del contexto del estudio el análisis se suele denominar experimento natural

  - Cuando esto ocurre, entonces se puede considerar que la asignación del tratamiento y la selección de individuos es aleatoria y, por tanto, las observaciones son independientes e idénticamente distribuidas

  - Uno de los problemas más importantes es la medición de los efectos de los tratamientos cuando C resulta de una decisión de

participación individual, de modo que los individuos pueden escoger voluntariamente si recibir el tratamiento y puede haber sesgo en el efecto del tratamiento

- Con el enfoque de experimento natural o cuasi-experimento, en el que se intenta replicar la asignación aleatoria, la aleatoriedad viene introducida por características del individuo en $\boldsymbol{x}$. En este caso, se asume que hay independencia condicional

- El modelo básico de selección anterior (en el modelo de Heckman, por ejemplo) se ha podido extender en muchas direcciones, y una de las aplicaciones más interesantes de este es su uso para medir los efectos de los tratamientos y de programas

  - En un modelo en el que se incluye la variable binaria del tratamiento C_i y las características $\boldsymbol{x_i}$ de los individuos, el coeficiente δ puede no estimar correctamente el efecto del tratamiento (aunque el resto de la regresión sea especificada correctamente)

$$y_i = \boldsymbol{x_i'\beta} + \delta C_i + \varepsilon_i$$

    - Esto ocurre porque puede haber problemas de selección muestral, en donde los individuos deciden o son más propensos a participar en el experimento o tratamiento por unas características concretas

    - De este modo, el efecto de los tratamientos se sobreestima o se subestima

  - Debido a este problema, se puede modelar como un modelo de dos ecuaciones en donde hay truncación incidental

$$C_i^* = \boldsymbol{w_i'\gamma} + u_i \qquad C_i^* = \begin{cases} 1 & if \;\; C_i^* > 0 \\ 0 & otherwise \end{cases}$$

$$E(y_i|C_i = 1, \boldsymbol{x_i}, \boldsymbol{w_i}) = \boldsymbol{x_i'\beta} + \boldsymbol{\delta} + E(\varepsilon_i|C_i = 1, \boldsymbol{x_i}, \boldsymbol{w_i}) =$$

$$= \boldsymbol{x_i'\beta} + \boldsymbol{\delta} + \rho\sigma_\varepsilon\lambda(\boldsymbol{w_i'\gamma})$$

  - Considerando un modelo de regresión para analizar los efectos del tratamiento en donde hay dos periodos (antes y después del tratamiento), es posible analizar los cambios a través del estimador de diferencias en diferencias

$$y_{it} = \boldsymbol{x_{it}'\beta} + \gamma C_i + \theta_t + u_i + \varepsilon_{it}$$

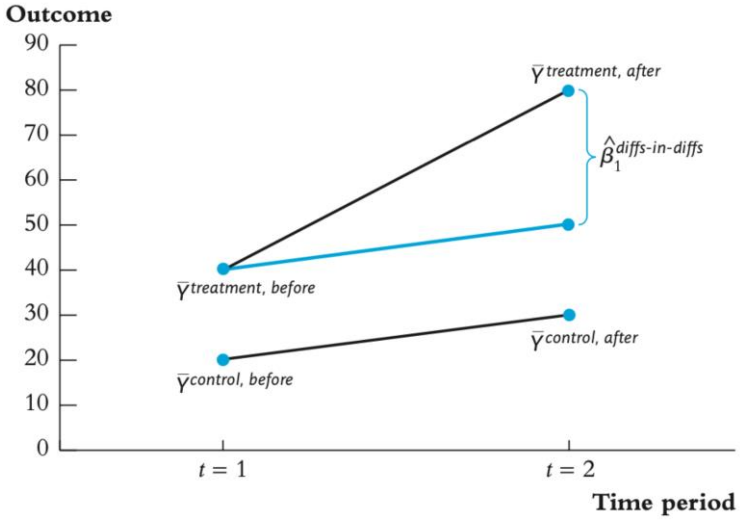
- En este caso, $\boldsymbol{x'_{it}}$ son las variables de control, C_i es la variable binaria para indicar el tratamiento, u_i son las características no observables del individuo y θ_t son los efectos temporales

- Considerando la presencia de variables de control en la regresión, la suposición de independencia del término de error y los regresores se cambia por la de independencia de media, en donde solo se pide que la variable del tratamiento sea independiente del error, pero no las variables de control

$$E(\varepsilon_i|C_i, \boldsymbol{x_{it}}) = E(\varepsilon_i|\boldsymbol{x_{it}})$$

- Como hay dos periodos, lo más natural es analizar los cambios cuando $\Delta C_i = 1$ para los individuos tratados y $\Delta C_i = 0$ para los no tratados (la diferencia entre ambos periodos). Como se puede ver, esta diferencia elimina los factores individuales no observados (son los mismos en el tiempo)

$$\Delta y_i = y_{i1} - y_{i0} = (\Delta \boldsymbol{x_{it}})'\boldsymbol{\beta} + \gamma \Delta C_i + (\theta_1 - \theta_0) + \Delta \varepsilon_{it}$$



- En la ausencia de variables de control $\boldsymbol{x'_{it}}$ de las características de los individuos, o asumiendo que no cambian en el tiempo, el estimador del efecto del tratamiento será el estimador de diferencias en diferencias

$$\gamma = E(\Delta y_i|\Delta C_i = 1) - E(\Delta y_i|C_i = 0) =$$

$$= [E(y_{i1}|C_i = 1) - E(y_{i0}|C_i = 1)] - [E(y_{i1}|C_i = 0) - E(y_{i0}|C_i = 0)]$$

$$\Rightarrow \hat{\gamma} = \left(\overline{\Delta y}|\Delta C_i = 1\right) - \left(\overline{\Delta y}|C_i = 0\right)$$

$$= [(\overline{y_1}|C = 1) - (\overline{y_0}|C = 1)] - [(\overline{y_1}|C = 0) - (\overline{y_0}|C = 0)]$$

- Aunque el problema se simplifica de manera considerable, el uso de diferencias hace que no se puede discernir lo que indujo al cambio (el tratamiento u otras características en $\boldsymbol{x_{it}}$). Además, las suposiciones son más fuertes de lo que gustaría un analista

- El contrafactual implícito es una observación de lo que la variable dependiente del individuo tratado sería si no fueran tratados (la diferencia para los no tratados), pero los individuos solo pueden estar en un estado, por lo que puede haber sesgo

    - El efecto medio del tratamiento en la población sería el ATE, que sería el efecto si se escogiera un individuo aleatoriamente de la población entera

$$ATE = E(y_1 - y_0)$$

    - No obstante, puede ser más interesante medir el efecto medio del tratamiento en los tratados, el cual sería el ATET. La dificultad con esto es que el contrafactual $E(y_0|C = 1)$ no se puede medir

$$ATET = E(y_1 - y_0|C = 1)$$

    - Si el tratamiento se asigna de manera totalmente aleatoria, entonces $E(y_j|C = 1) = E(y_j|C = 0) = E(y_j|C = j)$ para $j = 0,1$, por lo que el resultado potencial no depende del tratamiento y eso quiere decir que se puede medir realmente el efecto del tratamiento para aquellos que lo han recibido comparado con los que no

$$ATE = E(y_1|C = 1) - E(y_0|C = 0)$$

    - Sin embargo, si el tratamiento no se asigna de manera exógena o aleatoria, entonces es necesario utilizar otros enfoques como la función de control o variables instrumentales

- Cuando se intenta comparar la media o la esperanza de los resultados de dos poblaciones diferentes, estas se tienen que escalar para poder ser comparables

    - Para poder escalar las métricas, lo que se tiene que hacer es dividir la media por el error estándar de su respectiva población, de modo que la varianza es unitaria y permite comparar ambas métricas (que tendrán la misma variación)

$$\widehat{ATE} = \frac{\bar{y}_1}{\sqrt{S_1^2/n}} - \frac{\bar{y}_0}{\sqrt{S_0^2/n}}$$

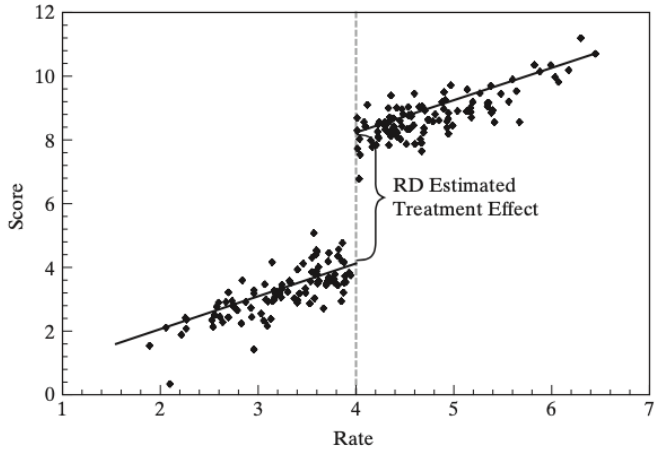
- Del mismo modo, para comparar estimaciones de los coeficientes de dos poblaciones diferentes, es necesario escalar los coeficientes y los errores estándar de los coeficientes estimados dividiendo entre el error estándar de la variables dependiente y

$$group\ effect = \frac{\hat{\beta}_1}{\sqrt{\frac{S_y^2}{n}}}$$

$$group\ effect\ std.error = \frac{S_{\widehat{\beta}_1}}{\sqrt{S_y^2/n}}$$

- Hay muchas situaciones en las que no hay posibilidad de asignar aleatoriamente individuos a los tratamientos, sobre todo en entornos que estudian las ciencias sociales. Con tal de superar este problema a la hora de la estimación, el diseño de regresión discontinua presenta una alternativa
  - Las condiciones bajo las que este enfoque es efectivo son que y sea una variable continua, que la variable dependiente varíe suavemente con una variable de asignación A y que el tratamiento se asigne de forma brusca en función del valor de A
    - Se puede considerar que la variable de asignación A es una característica del individuo que determina su participación en el tratamiento
    - Específicamente, la tercera condición indica que $C = 1_{A>A^*}$, donde A^* es un umbral fijo. Un diseño borroso o *fuzzy design*, en cambio, se basa en que $P(C = 1|A) = F(A)$, pero los problemas de identificación con este diseño son más complicados que con el diseño brusco
    - Por lo tanto, se asume que $y = f(A, C) + \varepsilon$. Este método, sin embargo, requiere que $E(\varepsilon|A, C) = E(\varepsilon|A)$, de modo que la variable de asignación es exógena al experimento
  - Por lo tanto, un modelo de regresión discontinua puede ser el siguiente, en donde α es el efecto del tratamiento a estimar:

$$y = f(A) + \alpha C + \varepsilon$$



- La especificación de $f(A)$ puede ser problemática, dado que asumir una función lineal cuando algo es muy general puede sesgar la estimación de α. Para estos casos se usan métodos no paramétricos, haciendo que el analista pueda utilizar observaciones que están más distantes del umbral

- La identificación del efecto de tratamiento comienza con la suposición de que $f(A)$ es continua en A^*, de modo que el efecto del tratamiento puede estimarse por la diferencia de los individuos que están cerca del valor de umbral A^*

$$\lim_{A \to A^*-} f(A) = \lim_{A \to A^*+} f(A) = f(A^*)$$

$$\Rightarrow \lim_{A \to A^*+} E(y|A) - \lim_{A \to A^*-} E(y|A) =$$

$$= f(A^*) + \alpha + \lim_{A \to A^*+} E(\varepsilon|A) - f(A^*) - \lim_{A \to A^*-} E(\varepsilon|A) = \alpha$$

- El efecto causal puede variar de un miembro de la población a otro, y este puede variar por variables observables $\boldsymbol{x}$ o por variables no observables (cuando se tiene una población heterogénea)

  - Cuando la población es heterogénea, se dan variaciones no observables en el efecto causal de cada individuo, de modo que cada individuo tiene su propio efecto causal

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \delta_i C_i + \varepsilon_i$$

    - Como el coeficiente varía de un individuo a otro, y los individuos se seleccionan aleatoriamente, esta es una variable aleatoria y no una constante como se asume anteriormente. El efecto medio del tratamiento será la esperanza del efecto causal para un individuo seleccionado aleatoriamente de la población a estudiar

$$ATE = E(y_1 - y_0) = E(\delta_i)$$

- Si el tratamiento se asigna aleatoriamente, entonces no está correlacionada con los resultados potenciales, y, por tanto, tampoco con δ_i ni con el error

- Cuando hay heterogeneidad en el efecto causal y el tratamiento se asigna aleatoriamente, el estimador de mínimos cuadrados ordinarios es consistente para estimar el efecto causal medio

$$\operatorname*{plim}_{p\to\infty} \delta_i = \frac{Cov(\boldsymbol{x}_i'\boldsymbol{\beta} + \delta_i C_i + \varepsilon_i, C_i)}{Var(C_i|\boldsymbol{X})} = \frac{Cov(\delta_i C_i, C_i)}{Var(C_i|\boldsymbol{X})} = \frac{E(\delta_i C_i C_i|\boldsymbol{X})}{Var(C_i|\boldsymbol{X})} =$$

$$= \frac{E(\delta_i|\boldsymbol{X})E(C_i C_i|\boldsymbol{X})}{Var(C_i|\boldsymbol{X})} = \frac{E(\delta_i|\boldsymbol{X})Var(C_i|\boldsymbol{X})}{Var(C_i|\boldsymbol{X})} = E(\delta_i|\boldsymbol{X})$$

- Cuando hay heterogeneidad en el efecto causal y el tratamiento se usa con un instrumento (válido), el estimador de mínimos cuadrados de dos fases o *TSLS* difiere, en general, del efecto medio del tratamiento

$$\operatorname*{plim}_{p\to\infty} \delta_i{}^{TSLS} = \frac{Cov(\boldsymbol{x}_i'\boldsymbol{\beta} + \delta_i(\boldsymbol{z}_i'\boldsymbol{\pi}_i + v_i) + \varepsilon_i, \boldsymbol{z}_i)}{Var(\boldsymbol{z}_i'\boldsymbol{\pi}_i + v_i|\boldsymbol{X})} = \frac{E(\delta_i\boldsymbol{\pi}_i|\boldsymbol{X})}{E(\boldsymbol{\pi}_i|\boldsymbol{X})} =$$

$$= E\left(\frac{\delta_i\boldsymbol{\pi}_i}{E(\boldsymbol{\pi}_i)}\middle|\boldsymbol{X}\right)$$

  - En este caso hay heterogeneidad en el efecto del instrumento sobre el regresor, de modo que los coeficientes $\boldsymbol{\pi}_i$ también son variables aleatorias e independientes del instrumento y de los términos de error

  - El estimador es una media ponderada de los efectos causales individuales, en donde los pesos son $\boldsymbol{\pi}_i/E(\boldsymbol{\pi}_i)$ y miden el grado relativo en el que el instrumento influencia la recepción del tratamiento para un individuo i. Este estimador se llama efecto medio del tratamiento local o ATEL, que enfatiza que se pesa más a individuos a los influencia más el instrumento

  - Las implicaciones que tiene esto es que los estimadores varíen en función de que variables instrumentales se escojan (dado que se obtienen estimadores ATEL diferentes), y, por tanto, la prueba con el *J-statistic* puede rechazar la hipótesis nula, aunque los instrumentos sean válidos

Los modelos para datos de panel

- Las bases de datos longitudinales o datos de panel son muy comunes en economía. Este tipo de datos son ricos en información y su análisis permite aprender sobre procesos económicos mientras se tiene en cuenta la heterogeneidad entre individuos y los efectos dinámicos que no aparecen en las secciones transversales. No obstante, su modelaje requiere especificaciones estocásticas complejas

    - Las bases de datos de panel o de datos longitudinales contienen secciones transversales grandes que se siguen en el tiempo, normalmente en pocos periodos

        - Los datos de panel están más orientados a análisis transversales en donde la heterogeneidad entre unidades es una parte integral del análisis, de modo que es posible analizar los cambios en el tiempo de las distribuciones transversales

        - El análisis de datos de panel es muy importante en econometría debido a que estas bases de datos proporcionan un ambiente rico para el desarrollo de técnicas de estimación y resultados teóricos

        - El uso de modelos enfocados a los datos de panel permite aislar los componentes de la varianza y poder controlar heterogeneidad no observada en diseños transversales (tales como la heterogeneidad que proviene del tiempo)

    - La ventaja fundamental de estos modelos de datos de panel sobre los modelos de datos transversales es que permiten mucha flexibilidad para modelar las diferencias de comportamiento entre individuos o grupos (la heterogeneidad individual). El marco básico para la discusión es el modelo de regresión siguiente:

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\alpha} + \varepsilon_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}$$

        - En esta regresión hay K regresores en $\boldsymbol{x}_{it}$ sin incluir un término constante, mientras que la heterogeneidad o efectos individuales se introducen mediante $c_i \equiv \boldsymbol{z}_i'\boldsymbol{\alpha}$, donde $\boldsymbol{z}_i$ contiene un término constante y un conjunto de variables específicas de los individuos o grupos que pueden ser observadas (raza, sexo, etc.) o no (características familiares, habilidad, etc.) que son constantes a lo largo del tiempo t

        - Si $\boldsymbol{z}_i$ se observa para todos los individuos o grupos, entonces el modelo entero se puede tratar como un modelo de regresión ordinario y se puede ajustar con mínimos cuadrados ordinarios. Pero si c_i no se puede observar, entonces hay más complicaciones (lo cual suele ser el caso)

- Este término se introduce para tener en cuenta de manera explícita el sesgo que se produciría por la heterogeneidad no observada a la hora de estimar los parámetros (permite controlarla)

- El objetivo del análisis es poder estimar de manera consistente y eficiente los efectos parciales de cada variable sobre y mientras se controla (se tiene en cuenta) la heterogeneidad

$$\boldsymbol{\beta} = \frac{\partial E(y_{it}|\boldsymbol{x_{it}})}{\partial \boldsymbol{x_{it}}}$$

  - Que eso sea posible depende de las suposiciones sobre los efectos que no se observan. Una de las suposiciones principales es la exogeneidad estricta de $\boldsymbol{x_{it}}$, de modo que el error no tiene correlación con las variables independientes en cualquier periodo pasado, presente y futuro)

$$E(\varepsilon_{it}|\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \dots) = 0$$

  - La segunda es la independencia media de c_i, de modo que si las variables no observadas no tienen correlación con las incluidas $\boldsymbol{x_{it}}$, entonces se pueden incluir en el error del modelo. Como es una suposición muy estricta, una más general sería que es una función de las variables independientes, pero es más complicada

$$E(c_i|\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \dots) = \alpha$$

$$\Rightarrow Cov(c_i, \boldsymbol{X_i}|\boldsymbol{X_i}) = Cov(E(c_i|\boldsymbol{X_i}), \boldsymbol{X_i}|\boldsymbol{X_i}) = Cov(\alpha, \boldsymbol{X_i}|\boldsymbol{X_i}) = 0$$

- Hay una variedad de modelos diferentes para datos de panel, pero los principales se pueden clasificar en cuatro:

  - En el modelo de regresión agrupada, $\boldsymbol{z_i}$ solo contiene un término constante, por lo que los estimadores de mínimos cuadrados ordinarios proporcionan estimadores eficientes y consistentes del parámetro α común y del vector $\boldsymbol{\beta}$

  - Si $\boldsymbol{z_i}$ no se observa, pero está correlacionado con $\boldsymbol{x_{it}}$, por lo que el estimador de mínimos cuadrados ordinarios de $\boldsymbol{\beta}$ está sesgado y es inconsistente por la omisión de variables. No obstante, si $\alpha_i \equiv c_i = \boldsymbol{z_i'\alpha}$, este incluye todos los efectos observables y especifica una media condicional estimable, y c_i será un término constante específico para cada grupo en el modelo, creando el modelo de efectos fijos (refiriéndose a la correlación entre c_i y $\boldsymbol{x_{it}}$)

$$y_{it} = \boldsymbol{x}'_{\boldsymbol{it}}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}$$

- Si la heterogeneidad de los individuos o grupos no observados no está correlacionada con las variables incluidas, se puede construir un modelo de regresión lineal con un error compuesto que puede ser estimado de manera consistente (pero no eficiente), dando lugar al modelo de efectos aleatorios. En este caso, u_i es un elemento aleatorio específico del grupo (similar a ε_{it} pero invariante en el tiempo)

$$y_{it} = \boldsymbol{x}'_{\boldsymbol{it}}\boldsymbol{\beta} + \boldsymbol{z}'_{\boldsymbol{i}}\boldsymbol{\alpha} + \varepsilon_{it} = \boldsymbol{x}'_{\boldsymbol{it}}\boldsymbol{\beta} + E(\boldsymbol{z}'_{\boldsymbol{i}}\boldsymbol{\alpha}) + [\boldsymbol{z}'_{\boldsymbol{i}}\boldsymbol{\alpha} - E(\boldsymbol{z}'_{\boldsymbol{i}}\boldsymbol{\alpha})] + \varepsilon_{it}$$

$$\Rightarrow\ y_{it} = \boldsymbol{x}'_{\boldsymbol{it}}\boldsymbol{\beta} + \alpha + u_i + \varepsilon_{it}$$

- El modelo de efectos aleatorios se puede ver como un modelo con coeficiente constante aleatorio, pero se puede extender esta idea a incluir otros coeficientes que varíen de manera aleatoria entre individuos o grupos. El modelo de parámetros aleatorios representa una extensión natural para aumentar la heterogeneidad, y $\boldsymbol{h_i}$ es un vector aleatorio que induce a la variación en los parámetros entre individuos o grupos

$$y_{it} = \boldsymbol{x}'_{\boldsymbol{it}}(\boldsymbol{\beta} + \boldsymbol{h_i}) + (\alpha + u_i) + \varepsilon_{it}$$

- Un aspecto importante del análisis de datos de panel es el balance del panel y su fijación

  - Si una base de datos de panel tiene n conjuntos de observaciones y cada individuo se observa por T periodos de tiempo, entonces el panel está balanceado

  - Si los individuos se observan por diferentes periodos de tiempo T_i, entonces el panel no está balanceado

  - Si la base de datos de panel estudia el mismo conjunto de individuos para toda la duración del estudio, el panel es fijo. Existen otros paneles en los que un conjunto de individuos se rota (queda fuera del estudio) en un periodo concreto, por lo que el panel de denomina rotativo

  - Los paneles también se pueden clasificar según la proporción de individuos con respecto a los periodos: si T es mucho mayor a n, el panel se denomina panel largo o *long panel*, mientras que en el caso contrario (con T mucho menor a n) se denomina panel corto o *short panel*

- Las propiedades asintóticas de los estimadores en el modelo de regresión múltiple se establecieron bajo las suposiciones anteriormente vistas (linealidad, rango máximo, exogeneidad de variables independientes, homoscedasticidad y no autocorrelación y observaciones i.i.d.)

    - No obstante, hay veces que al tratar con datos de panel se tienen hacer excepciones, dado que la muestra consiste de múltiples observaciones en cada una de las unidades observacionales, por lo que las $\boldsymbol{x}$ pueden estar correlacionadas entre observaciones en el tiempo (al menos dentro de cada unidad, aunque podría haber unidades correlacionadas también)

    - Asumiendo que los datos constan de un número fijo de observaciones T en un conjunto de n individuos, el número total de filas en $\boldsymbol{X}$ será $N = nT$, y la matriz $\overline{\boldsymbol{Q}}_{\boldsymbol{n}} \equiv \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Q}_{\boldsymbol{i}}$ se puede expresar de la siguiente manera:

$$\overline{\boldsymbol{Q}}_{\boldsymbol{n}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Q}_{\boldsymbol{i}} = \frac{1}{n}\sum_{i} \frac{1}{T}\sum_{t \in i} \boldsymbol{Q}_{\boldsymbol{it}} = \frac{1}{n}\sum_{i=1}^{n} \overline{\boldsymbol{Q}}_{\boldsymbol{i}}$$

$$\overline{\boldsymbol{Q}}_{\boldsymbol{i}} = average\ \boldsymbol{Q}_{\boldsymbol{it}}\ for\ individual$$

    - Las condiciones que se necesitan para establecer convergencia aplican con respecto al número total de unidades observacionales que se consideren (en donde cada grupo puede ser una sola observación o un conjunto de observaciones)

- Un modelo más generalizado para el marco de las regresiones de datos de panel, es el siguiente:

$$y_{it} = \boldsymbol{x}_{\boldsymbol{it}}'\boldsymbol{\beta}_{\boldsymbol{it}} + u_{it} \quad where \quad u_{it} = \alpha_i + \delta_t + \varepsilon_{it}$$

    - En este modelo, el error u_{it} se puede dividir en perturbaciones que provienen de la dimensión individual, de la dimensión temporal, y de ambas a la vez

    - Por lo tanto, las suposiciones de homoscedasticidad y de no autocorrelación no son adecuadas en este caso, aunque lo que normalmente se asume para este modelo son las siguientes:

$$E(u_{it}|\boldsymbol{x}_{\boldsymbol{it}}) = 0 \qquad Var(u_{it}) = \sigma_{it}^2 \qquad Cov(u_{it}, u_{js}) = \sigma_{ij,ts}$$

    - Los parámetros pueden ser estimados por mínimos cuadrados ordinarios o por mínimos cuadrados generalizados dadas estas suposiciones, aunque son muy simplistas

- Muchas veces se asume que $\delta_t = 0$, aunque no siempre, dado que hay modelos que consideran los efectos temporales

- El análisis se comienza asumiendo la versión más simple del modelo general de datos de panel, el modelo agrupado o *pooled model*. En esta forma, si las suposiciones del modelo de regresión múltiple se cumplen, entonces no es necesario otros estimadores más que el de mínimos cuadrados ordinarios

$$y_{it} = \alpha + \boldsymbol{x}_{\boldsymbol{it}}'\boldsymbol{\beta} + \varepsilon_{it} \quad for \;\; i = 1, \dots, n \;\; \& \;\; t = 1, \dots, T$$

$$E\left(\varepsilon_{it} \middle| \boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \dots, \boldsymbol{x_{iT_i}}\right) = 0$$

$$Var\left(\varepsilon_{it} \middle| \boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \dots, \boldsymbol{x_{iT_i}}\right) = \sigma_\varepsilon^2$$

$$Cov\left(\varepsilon_{it}, \varepsilon_{is} \middle| \boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \dots, \boldsymbol{x_{iT_i}}\right) = 0 \;\; if \;\; i \neq j \;\; or \; t \neq s$$

  - Las suposiciones subyacentes de la estimación por mínimos cuadrados ordinarios que también se asumen en este modelo, no obstante, no suelen cumplirse como se asume en el modelo agrupado

    - En este caso, el modelo muestra que la heterogeneidad no observada es la misma para todos los individuos (se trata como el término constante de la regresión)

    - La heterogeneidad no observada induce a la autocorrelación de los errores en el tiempo, haciendo que el estimador asintótico de mínimos cuadrados de la varianza esté sesgado

    - Esto se puede ver de manera clara a través del modelo de efectos aleatorios que se verá más adelante, en donde el término de error para los grupos u_i está correlacionado con $\boldsymbol{x_{it}}$ y por tanto introduce sesgo en los estimadores

$$y_{it} = \alpha + \boldsymbol{x}_{\boldsymbol{it}}'\boldsymbol{\beta} + \varepsilon_{it} + \left(c_i - E(c_i|\boldsymbol{X_i})\right) = \alpha + \boldsymbol{x}_{\boldsymbol{it}}'\boldsymbol{\beta} + \varepsilon_{it} + u_i =$$

$$= \alpha + \boldsymbol{x}_{\boldsymbol{it}}'\boldsymbol{\beta} + w_{it}$$

$$\Rightarrow \; E(w_{it}w_{is}) = Cov(w_{it}, w_{is}) = Cov(\varepsilon_{it} + u_i, \varepsilon_{is} + u_i) =$$

$$= Cov(u_i, u_i) = \sigma_u^2 \;\; for \; t \neq s$$

  - Suponiendo un modelo aún más general en las que se juntan todas las observaciones T_i para un individuo i en una sola ecuación y que el vector de perturbaciones $\boldsymbol{w_i}$ incluye ε_{it} y los componentes omitidos, se dan las siguientes igualdades:

$$\boldsymbol{y_i} = \boldsymbol{X_i\beta} + \boldsymbol{w_i} \qquad Var(\boldsymbol{w_i}|\boldsymbol{X_i}) = \boldsymbol{\sigma_\varepsilon^2 I}_{T_i} + \boldsymbol{\Sigma_i} = \boldsymbol{\Omega_i}$$

- El estimador de mínimos cuadrados ordinarios $\boldsymbol{\beta}$, bajo autocorrelación, será el siguiente:

$$\boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} = \left[\sum_{i=1}^{n} \boldsymbol{X_i'X_i}\right]^{-1} \sum_{i=1}^{n} \boldsymbol{X_i'y_i} =$$

$$= \left[\sum_{i=1}^{n} \boldsymbol{X_i'X_i}\right]^{-1} \sum_{i=1}^{n} \boldsymbol{X_i'}(\boldsymbol{X_i\beta} + \boldsymbol{w_i}) = \boldsymbol{\beta} + \left[\sum_{i=1}^{n} \boldsymbol{X_i'X_i}\right]^{-1} \sum_{i=1}^{n} \boldsymbol{X_i'w_i}$$

- La consistencia se puede establecer de la misma manera que anteriormente, y la verdadera matriz de covarianzas asintótica tomaría la forma que se vio para el modelo de regresión generalizado

$$Asy.Var(\boldsymbol{b}) =$$

$$= \frac{1}{n}\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X_i'X_i}\right]^{-1} \text{plim}\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X_i'w_iw_i'X_i}\right] \text{plim}\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X_i'X_i}\right]^{-1} =$$

$$= \frac{1}{n}\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X_i'X_i}\right]^{-1} \text{plim}\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X_i'\Omega_iX_i}\right] \text{plim}\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X_i'X_i}\right]^{-1}$$

- La matriz central tiene que ser estimada, y se puede estimar de la misma manera que con el estimador de White (la lógica de este también aplica para este caso, aunque no sea exactamente lo mismo), en donde $\widehat{\boldsymbol{w}}_{\boldsymbol{i}}'$ es el vector de T_i residuos para el individuo i

$$Est.Asy.Var(\boldsymbol{b}) =$$

$$= \frac{1}{n}\text{plim}\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X_i'X_i}\right]^{-1} \text{plim}\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X_i'}\widehat{\boldsymbol{w}}_{\boldsymbol{i}}\widehat{\boldsymbol{w}}_{\boldsymbol{i}}^{'}\boldsymbol{X_i}\right] \text{plim}\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X_i'X_i}\right]^{-1}$$

- No obstante, el problema más importante no es la heteroscedasticidad sino la autocorrelación entre observaciones, de modo que el estimador de White no es una solución al problema de inferencia

- Muchos estudios se realizan sacando muestras de clústeres, ya que de otro modo se pueden surgir naturalmente efectos similares a los efectos aleatorios en los datos de panel

- Los efectos de cada clúster dentro de la muestra pueden inducir a la correlación entre observaciones similares a los efectos aleatorios o fijos que se han identificado anteriormente, lo cual, si se ignoran, pueden llevar a errores de inferencia graves

- Para un modelo de dos niveles (en donde hay clústeres dentro de la muestra), un enfoque natural que se puede llevar a cabo es utilizar el enfoque robusto de efectos agrupados visto anteriormente

- El estimador de clústeres para para un modelo de un solo nivel es muy parecido al del modelo de regresión de efectos agrupados, pero hay una diferencia en el proceso de generación de datos, dado que, en el primer ajuste, los individuos en el grupo generalmente se observan una vez, y su asociación (los efectos comunes) es menos definida

- Para poder tener en cuenta los efectos en muestras pequeñas cuando el número de clústeres es un porcentaje significativo de un total finito, se utiliza una versión refinada del estimador anteriormente visto con una corrección de los grados de libertad

$$Est.Asy.Var(\boldsymbol{b}) =$$

$$= \text{plim}\left[\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{X}_g\right]^{-1}\text{plim}\left[\frac{G}{G-1}\sum_{g=1}^{G}\left(\sum_{i=1}^{n_g}\boldsymbol{x}_{ig}\widehat{\boldsymbol{w}}_{ig}\right)\left(\sum_{i=1}^{n_g}\boldsymbol{x}_{ig}\widehat{\boldsymbol{w}}_{ig}\right)\right]\text{plim}\left[\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{X}_g\right]^{-1} =$$

$$= \text{plim}\left[\sum_{g=1}^{G}\boldsymbol{X}_{\boldsymbol{g}}'\boldsymbol{X}_{\boldsymbol{g}}\right]^{-1}\text{plim}\left[\frac{G}{G-1}\sum_{g=1}^{G}\boldsymbol{X}_{\boldsymbol{i}}'\widehat{\boldsymbol{w}}_{\boldsymbol{i}}\widehat{\boldsymbol{w}}_{\boldsymbol{i}}'\boldsymbol{X}_{\boldsymbol{i}}\right]\text{plim}\left[\sum_{g=1}^{G}\boldsymbol{X}_{\boldsymbol{g}}'\boldsymbol{X}_{\boldsymbol{g}}\right]^{-1}$$

$$G = n^{\underline{o}}\ clusters\ in\ sample \qquad n_g = n^{\underline{o}}\ indiv.in\ g$$

- Si el número de clústeres es pequeño, entonces la corrección de clústeres utilizada después de estimar un modelo de regresión agrupada por mínimos cuadrados ordinarios no es muy útil

- El modelo de regresión agrupado se puede estimar utilizando las medias muestrales de los datos. El modelo de regresión implícito se obtiene multiplicando cada grupo por $(1/T)\boldsymbol{i}'$, en donde $\boldsymbol{i}$ es en un vector columna de unos

$$(1/T)\boldsymbol{i}'\boldsymbol{y}_{\boldsymbol{i}} = (1/T)\boldsymbol{i}'\boldsymbol{x}_{\boldsymbol{i}}\boldsymbol{\beta} + (1/T)\boldsymbol{i}'\boldsymbol{w}_{\boldsymbol{i}} \quad \Rightarrow \quad \bar{y}_i = \bar{\boldsymbol{x}}_{\boldsymbol{i}}\boldsymbol{\beta} + \bar{w}_i$$

- En el modelo de regresión lineal transformado, las perturbaciones siguen teniendo una media condicional nula, pero varianzas heteroscedásticas. Con $\boldsymbol{\Omega_i}$ sin especificar, la regresión es heteroscedástica y se puede utilizar el estimador de White, pero si las suposiciones clásicas no se mantienen, entonces no se resuelve el problema

- Utilizar la media de grupos no resuelve el problema, pero la pérdida de información que ocurre al hacer la media es relativamente pequeña (aunque los datos desagregados sean mejores) y se permite eliminar la dimensión temporal en el modelo

- Considerando un modelo general en el que c_i es el efecto latente (variable aleatoria), si la suposición de independencia media $E(c_i|\boldsymbol{X_i}) = \alpha$ no se cumple, entonces el efecto latente también se transmite a las medias grupales

$$E(c_i|\boldsymbol{X_i}) = h(\boldsymbol{X_i}) = \overline{\boldsymbol{x}}_{\boldsymbol{i}}\boldsymbol{\gamma}$$

$$y_{it} = \boldsymbol{x}'_{\boldsymbol{it}}\boldsymbol{\beta} + c_i + \varepsilon_{it} = \boldsymbol{x}'_{\boldsymbol{it}}\boldsymbol{\beta} + \overline{\boldsymbol{x}}_{\boldsymbol{i}}\boldsymbol{\gamma} + [\varepsilon_{it} + c_i - \overline{\boldsymbol{x}}_{\boldsymbol{i}}\boldsymbol{\gamma}] =$$

$$= \boldsymbol{x}'_{\boldsymbol{it}}\boldsymbol{\beta} + \overline{\boldsymbol{x}}_{\boldsymbol{i}}\boldsymbol{\gamma} + u_{it}$$

$$\Rightarrow \bar{y}_i = \overline{\boldsymbol{x}}_{\boldsymbol{i}}\boldsymbol{\beta} + \overline{\boldsymbol{x}}_{\boldsymbol{i}}\boldsymbol{\gamma} + \bar{u}_i = \overline{\boldsymbol{x}}_{\boldsymbol{i}}(\boldsymbol{\beta} + \boldsymbol{\gamma}) + \bar{u}_i$$

- En ese caso, el estimador de medias grupales estima $\boldsymbol{\beta} + \boldsymbol{\gamma}$ y no $\boldsymbol{\beta}$, por lo que se agrupan los efectos observados y no observados en el mismo estimador. Si, en otro caso, hubiera error de medición en los regresores, este error se disiparía en la media y el estimador seguiría siendo consistente (aunque el de mínimos cuadrados no)

- El modelo de regresión agrupada se puede estimar de manera consistente (no eficiente) con el estimador de mínimos cuadrados ordinarios. Si se consideran las matrices de sumas de cuadrados y productos cruzados para cada formulación equivalente del modelo, se obtienen las siguientes identidades:

  - En el modelo original, los momentos acumularán la variación en las medias totales $\bar{\bar{y}}$ y $\overline{\overline{\boldsymbol{x}}}$ y se usarían las siguientes sumas en el estimador:

$$y_{it} = \alpha + \boldsymbol{x}'_{\boldsymbol{it}}\boldsymbol{\beta} + \varepsilon_{it}$$

$$\boldsymbol{S_{xx}^{total}} = \sum_{i=1}^{n}\sum_{t=1}^{T}(\boldsymbol{x_{it}} - \bar{\bar{\boldsymbol{x}}})(\boldsymbol{x_{it}} - \bar{\bar{\boldsymbol{x}}})'$$

$$\boldsymbol{S_{xy}^{total}} = \sum_{i=1}^{n}\sum_{t=1}^{T}(\boldsymbol{x_{it}} - \bar{\bar{\boldsymbol{x}}})(y_{it} - \bar{\bar{y}})$$

- En el modelo en términos de la media de grupos, la media de la media de los grupos es la media total, por lo que las matrices de momentos son sumas de cuadrados entre grupos (la variación de las medias de los grupos con respecto a las medias totales) y productos cruzados

$$\bar{y}_i = \alpha + \bar{\boldsymbol{x}}_{\boldsymbol{i}}\boldsymbol{\beta} + \bar{\varepsilon}_i$$

$$\boldsymbol{S_{xx}^{between}} = T\sum_{i=1}^{n}(\bar{\boldsymbol{x}}_{\boldsymbol{i}} - \bar{\bar{\boldsymbol{x}}})(\bar{\boldsymbol{x}}_{\boldsymbol{i}} - \bar{\bar{\boldsymbol{x}}})'$$

$$\boldsymbol{S_{xy}^{between}} = \sum_{i=1}^{n}T(\bar{\boldsymbol{x}}_{\boldsymbol{i}} - \bar{\bar{\boldsymbol{x}}})(\bar{y}_i - \bar{\bar{y}})$$

- En el modelo en términos de desviaciones de la media grupal, como los datos ya están en desviaciones, las medias de estas son nulas y las matrices de momentos son sumas de cuadrados dentro de los grupos (la variación de las observaciones con respecto a las medias grupales) y productos cruzados

$$y_{it} - \bar{y}_i = (\boldsymbol{x_{it}} - \bar{\boldsymbol{x}}_{\boldsymbol{i}})'\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_i$$

$$\boldsymbol{S_{xx}^{within}} = \sum_{i=1}^{n}\sum_{t=1}^{T}(\boldsymbol{x_{it}} - \bar{\boldsymbol{x}}_{\boldsymbol{i}})(\boldsymbol{x_{it}} - \bar{\boldsymbol{x}}_{\boldsymbol{i}})'$$

$$\boldsymbol{S_{xy}^{within}} = \sum_{i=1}^{n}\sum_{t=1}^{T}(\boldsymbol{x_{it}} - \bar{\boldsymbol{x}}_{\boldsymbol{i}})(y_{it} - \bar{y}_i)'$$

- Se puede verificar que la suma de cuadrados total y el producto cruzado total es la suma de aquellos entre grupos y dentro de los grupos

$$\boldsymbol{S_{xx}^{total}} = \boldsymbol{S_{xx}^{within}} + \boldsymbol{S_{xx}^{between}}$$

$$\boldsymbol{S_{xy}^{total}} = \boldsymbol{S_{xy}^{within}} + \boldsymbol{S_{xy}^{between}}$$

- Por lo tanto, hay tres posibles estimadores de mínimos cuadrados para el vector $\boldsymbol{\beta}$ del modelo agrupado, que corresponden a la descomposición de $\boldsymbol{S_{xx}^{total}}$ y $\boldsymbol{S_{xy}^{total}}$

  - Para el modelo original, el estimador de mínimos cuadrados es el original. Este es consistente y no está sesgado, pero no es el más eficiente

$$\boldsymbol{b^{total} = \left[S_{xx}^{total}\right]^{-1} S_{xy}^{total} = \left[S_{xx}^{between} + S_{xx}^{within}\right]^{-1}\left[S_{xy}^{between} + S_{xy}^{within}\right]}$$

  - Para el modelo en términos de medias grupales, el estimador de mínimos cuadrados es el estimador entre grupos, el cual es el estimador de medias grupales visto anteriormente. Este es consistente y no está sesgado cuando $nT \to \infty$, pero, aunque tampoco sea eficiente, tiene menor varianza que el original

$$\boldsymbol{b^{between} = \left[S_{xx}^{between}\right]^{-1}\left[S_{xy}^{between}\right]}$$

  - Para el modelo en términos de medias grupales, el estimador de mínimos cuadrados es el estimador dentro de grupos. Este es consistente y no está sesgado cuando $nT \to \infty$, y además tiene menos varianza que los dos estimadores anteriores (aunque no es el más eficiente)

$$\boldsymbol{b^{within} = \left[S_{xx}^{within}\right]^{-1}\left[S_{xy}^{within}\right]}$$

  - A partir de estos estimadores, es posible encontrar una expresión equivalente para los productos cruzados y así, encontrar una expresión equivalente de $\boldsymbol{b^{total}}$ en términos de matrices

$$\boldsymbol{S_{xy}^{between} = S_{xx}^{between} b^{between}}$$

$$\boldsymbol{S_{xy}^{within} = S_{xx}^{within} b^{within}}$$

$$\boldsymbol{\Rightarrow b^{total} = \left[S_{xx}^{between} + S_{xx}^{within}\right]^{-1}\left[S_{xx}^{between} b^{between} + S_{xx}^{within} b^{within}\right]}$$

$$\boldsymbol{\Rightarrow\ b^{total} = F^{between} b^{between} + F^{within} b^{bwithin}}$$

$$\boldsymbol{where\ \ F^{between} \equiv \left[S_{xx}^{between} + S_{xx}^{within}\right]^{-1} S_{xx}^{between} = I - F^{within}}$$

  - Cuando $T = 2$, se puede demostrar que el estimador de mínimos cuadrados ordinarios, el estimador entre grupos y el de primeras diferencias coinciden entre sí

$$\boldsymbol{b^{within}} = \frac{\sum_{i=1}^{n}\left(\frac{\boldsymbol{x_{i2}+x_{i1}}}{2}\right)\left(\frac{y_{i2}+y_{i1}}{2}\right)}{\sum_{i=1}^{n}\left(\frac{\boldsymbol{x_{i2}+x_{i1}}}{2}\right)\left(\frac{\boldsymbol{x_{i2}+x_{i1}}}{2}\right)'} = \frac{\sum_{i=1}^{n}(y_{i2}+y_{i1})}{\sum_{i=1}^{n}(\boldsymbol{x_{i2}+x_{i1}})}$$

$$\boldsymbol{b} = \frac{\sum_{i=1}^{n}\left(\frac{\boldsymbol{x_{i2}+x_{i1}}}{2}\right)\left(\frac{y_{i2}+y_{i1}}{2}\right)}{\sum_{i=1}^{n}\left(\frac{\boldsymbol{x_{i2}+x_{i1}}}{2}\right)\left(\frac{\boldsymbol{x_{i2}+x_{i1}}}{2}\right)'} = \frac{\sum_{i=1}^{n}(y_{i2}+y_{i1})}{\sum_{i=1}^{n}(\boldsymbol{x_{i2}+x_{i1}})}$$

$$\boldsymbol{b^{FD}} = \frac{\sum_{i=1}^{n}\left(\frac{\boldsymbol{x_{i2}+x_{i1}}}{2}\right)\left(\frac{y_{i2}+y_{i1}}{2}\right)}{\sum_{i=1}^{n}\left(\frac{\boldsymbol{x_{i2}+x_{i1}}}{2}\right)\left(\frac{\boldsymbol{x_{i2}+x_{i1}}}{2}\right)'} = \frac{\sum_{i=1}^{n}(y_{i2}+y_{i1})}{\sum_{i=1}^{n}(\boldsymbol{x_{i2}+x_{i1}})}$$

$$T = 2 \quad \Rightarrow \quad \boldsymbol{b = b^{within} = b^{FD}}$$

- El modelo de efectos fijos nace de la suposición sobre los efectos no observados en el modelo general c_i están correlacionados con las variables incluidas $\boldsymbol{x_{it}}$, de modo que añadiendo este término c_i a la regresión se puede compensar por el sesgo introducido por esta correlación

$$y_{it} = \boldsymbol{x_{it}'\beta} + c_i + \varepsilon_{it} = \boldsymbol{x_{it}'\beta} + h(\boldsymbol{X_i}) + \varepsilon_{it} + [c_i - h(\boldsymbol{X_i})]$$

$$\Rightarrow\ y_{it} = \boldsymbol{x_{it}'\beta} + \alpha_i + \varepsilon_{it} + [c_i - h(\boldsymbol{X_i})]$$

  - Se supone que c_i es una variable aleatoria con $E(c_i|\boldsymbol{X_i}) = h(\boldsymbol{X_i})$, de modo que $c_i - h(\boldsymbol{X_i})$ no tiene correlación con $\boldsymbol{X_i}$ (por los argumentos vistos anteriormente) y se puede absorber en el término de error ε_{it}

$$\Rightarrow E(y_{it}|\boldsymbol{x_{it}}) = E(\boldsymbol{x_{it}'\beta} + h(\boldsymbol{X_i}) + \varepsilon_{it} + [c_i - h(\boldsymbol{X_i})]|\boldsymbol{x_{it}}) = \boldsymbol{x_{it}'\beta} + \alpha_i$$

$$\Rightarrow\ Cov(c_i - h(\boldsymbol{X_i}), \boldsymbol{X_i}|\boldsymbol{X_i}) = Cov(E(c_i|\boldsymbol{X_i}) - h(\boldsymbol{X_i}), \boldsymbol{X_i}|\boldsymbol{X_i}) =$$

$$= Cov(h(\boldsymbol{X_i}) - h(\boldsymbol{X_i}), \boldsymbol{X_i}|\boldsymbol{X_i}) = Cov(0, \boldsymbol{X_i}|\boldsymbol{X_i}) = 0$$

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

    - En verdad, los efectos fijos sí que podrían estar correlacionados con los regresores, pero se requiere que no para la consistencia de los estimadores en este modelo. De estar correlacionados, los estimadores no serán consistentes

    - Otra suposición que se hace que es que $Var(c_i|\boldsymbol{X_i})$ es constante, por lo que con ambas suposiciones se obtiene un modelo de regresión lineal múltiple

$$Var(c_i|\boldsymbol{X_i}) = \sigma_c^2$$

- Qué $E(c_i|\boldsymbol{X_i}) = h(\boldsymbol{X_i})$ significa que los efectos latentes son fijos en el tiempo (no dependen del tiempo), y la formulación implica que las diferencias entre grupos se pueden capturar en las diferencias en el término constante α_i (comparando la de un grupo con otro). Cada α_i es un parámetro desconocido que se tiene que estimar

- Una de las mayores desventajas de la regresión de efectos fijos es que cualquier variable $\boldsymbol{x_{it}}$ que no varíe en el tiempo será absorbida en α_i y los coeficientes de estas variables no se podrían estimar con los dos estimadores vistos anteriormente

$$\overline{\boldsymbol{x}}_{\boldsymbol{i}} = \frac{\sum_{t=1}^{T} \boldsymbol{x_i}}{T} = \boldsymbol{x_i} \quad for \;\; i = 1,2, \dots, N$$

- Las suposiciones hechas tienen sentido siempre que la relación N/T no sea muy grande, ya que de otro modo el modelo tendría pocos grados de libertad

$$Degrees\ of\ Freedom = N(T-1) - K$$

- Siendo $\boldsymbol{y_i}$ y $\boldsymbol{X_i}$ las T observaciones para la unidad observacional i, $\boldsymbol{i}$ un vector columna $T \times 1$ y $\boldsymbol{\varepsilon_i}$ un vector columna $T \times 1$ de perturbaciones, el modelo se puede expresa de la siguiente manera:

$$\boldsymbol{y_i} = \boldsymbol{X_i\beta} + \boldsymbol{i}\alpha_i + \boldsymbol{\varepsilon_i}$$

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{i} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{i} & \cdots & \mathbf{0} \\ & & \vdots & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{i} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix}$$

- Este se puede expresar de manera equivalente creando una matriz $\boldsymbol{D} \equiv [\boldsymbol{d_1}|\boldsymbol{d_2}| \dots |\boldsymbol{d_n}]$ de tamaño $nT \times n$, donde $\boldsymbol{d_i}$ es una variable binaria que indica la unidad i. Juntando las nT filas, se obtiene la siguiente expresión, denominado modelo de mínimos cuadrados de variable binaria (LSDV):

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{D\alpha} + \boldsymbol{\varepsilon}$$

- Como el modelo es uno de regresión múltiple común, se puede utilizar el estimador de mínimos cuadrados ordinarios, que será equivalente al estimador dentro de grupos (en donde se les resta la media temporal a los regresores):

$$\boldsymbol{b^{LSDV}} = [\boldsymbol{X'M_DX}]^{-1}[\boldsymbol{X'M_Dy}] = \boldsymbol{b^{within}}$$

$$where\ \ \boldsymbol{M_D} = \boldsymbol{I} - \boldsymbol{D}(\boldsymbol{D'D})^{-1}\boldsymbol{D'}$$

- Igual que antes, si $T = 2$, entonces también se pueden estimar los coeficientes por primeras diferencias y los estimadores de mínimos cuadrados y de dentro de grupos son equivalentes

$$\boldsymbol{b^{LSDV}} = \boldsymbol{b^{within}} = \boldsymbol{b^{FD}}$$

- El estimador entre grupos no se puede utilizar aquí debido a que los efectos fijos se verían absorbidos a utilizar el modelo de medias temporales de las variables

$$\bar{\alpha}_i = \frac{\sum_{t=1}^{T} \alpha_i}{T} = \alpha_i \quad for \ \ i = 1,2, \dots, N$$

- Los coeficientes de las variables binarias $\boldsymbol{d_i}$ se pueden recuperar a partir de la ecuación normal de la regresión parcial, de modo que se puede obtener el coeficiente a_i^{LSDV} para cada i, siendo un estimador para a_i^{LSDV} consistente cuando $T \to \infty$

$$\boldsymbol{D'Da} + \boldsymbol{D'Xb^{LSDV}} = \boldsymbol{D'y} \ \Rightarrow \boldsymbol{a} = [\boldsymbol{D'D}]^{-1}\boldsymbol{D'}(\boldsymbol{y} - \boldsymbol{Xb^{LSDV}})$$

$$\Rightarrow \ a_i = \bar{y}_i - \bar{\boldsymbol{x}}_{\boldsymbol{i}}'\boldsymbol{b^{LSDV}}$$

- El estimador asintótico de la matriz de covarianza para $\boldsymbol{b^{LSDV}}$ utiliza el segundo momento de la matriz con $\boldsymbol{x}$ que se expresan como desviaciones de sus respectivas medias grupales

$$Est.Asy.Var(\boldsymbol{b^{LSDV}}) = s^2[\boldsymbol{X'M_DX}]^{-1} = s^2\left[\boldsymbol{S_{xx}^{within}}\right]^{-1}$$

$$s^2 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \boldsymbol{x_{it}'b^{LSDV}} - a_i)^2}{nT - n - K} =$$

$$= \frac{(\boldsymbol{M_Dy} - \boldsymbol{M_DXb^{LSDV}})'(\boldsymbol{M_Dy} - \boldsymbol{M_DXb^{LSDV}})}{nT - n - K}$$

- El numerador en s^2 es exactamente la suma de residuos cuadrados utilizando los coeficientes de mínimos cuadrados ordinarios y los datos en forma de desviaciones de la media grupal

$$e_{it} = y_{it} - \boldsymbol{x_{it}'b^{LSDV}} - a_i = y_{it} - \boldsymbol{x_{it}'b^{LSDV}} - \left(\bar{y}_i - \bar{\boldsymbol{x}}_{\boldsymbol{i}}'\boldsymbol{b^{LSDV}}\right)$$

$$= (y_{it} - \bar{y}_i) + (\boldsymbol{x_{it}} - \bar{\boldsymbol{x}}_{\boldsymbol{i}})'\boldsymbol{b^{LSDV}}$$

- Para los efectos individuales, el estimador de la varianza asintótica de a_i es el siguiente:

$$Asy.Var(a_i) = \frac{\sigma_\varepsilon^2}{T} + \bar{\boldsymbol{x}}_{\boldsymbol{i}}'[Asy.Var(\boldsymbol{b}^{\boldsymbol{LSDV}})]\bar{\boldsymbol{x}}_{\boldsymbol{i}}$$

- Es posible utilizar un contraste t para contrastar $a_i = 0$, pero esto es solo para un grupo específico i y no sirve en este contexto de regresión. Por lo tanto, si uno está interesado en las diferencias entre grupos, se puede contrastar la hipótesis de que todos los términos constantes son iguales con un estadístico F

$$F(n-1, nT-n-K) = \frac{(R_{LSDV}^2 - R_{Pooled}^2)/(n-1)}{(1-R_{LSDV}^2)/(nT-n-K)}$$

  - En este caso se compara el ajuste del modelo estimado con variables binarias y del modelo agrupado, dado que el modelo agrupado solo tiene un término constante (la heterogeneidad no observada) que es igual para todos los grupos

  - Si se rechaza la hipótesis nula, entonces el modelo de efectos fijos es mejor que el modelo agrupado para modelar los datos (tiene mejor ajuste) porque la heterogeneidad no observada no es la misma entre individuos. De no rechazarse, la conclusión es la contraria

- El enfoque de regresión de mínimos cuadrados de variable binaria se puede extender para incluir efectos específicos del tiempo, de modo que el modelo se puede extender del siguiente modo:

$$y_{it} = \boldsymbol{x}_{\boldsymbol{it}}'\boldsymbol{\beta} + \alpha_i + \delta_t + \varepsilon_{it}$$

  - Este modelo se obtiene del anterior incluyendo $T-1$ variables binarias adicionales (a través del término δ_t) como al construir la matriz $\boldsymbol{D}$ anterior (una no se incluye para evitar la trampa)

  - No obstante, hay una asimetría en la formulación, dado que cada efecto de grupo tiene un intercepto específico mientras que los efectos temporales son contrastes (comparaciones hechas respecto al periodo base excluido), por lo que se puede formular el siguiente modelo simétrico en donde se incluyen los n y T efectos (coeficientes de las variables binarias) de manera restringida:

$$y_{it} = \boldsymbol{x}_{\boldsymbol{it}}'\boldsymbol{\beta} + \mu + \alpha_i + \delta_t + \varepsilon_{it} \quad s.t. \quad \sum_i \alpha_i = \sum_i \delta_t = 0$$

- Los estimadores por mínimos cuadrados ordinarios en este modelo se obtienen con una regresión de y_{it}^* sobre $\boldsymbol{x_{it}^*}$, y los coeficientes de las variables binarias y del coeficiente constante se pueden recuperar de las ecuaciones normales

$$y_{it}^* = y_{it} - \bar{y}_i - \bar{y}_t + \bar{\bar{y}} \qquad \boldsymbol{x_{it}^*} = \boldsymbol{x_{it}} - \boldsymbol{\bar{x}_i} - \boldsymbol{\bar{x}_t} + \boldsymbol{\bar{\bar{x}}}$$

$$\hat{\mu} = m = \bar{\bar{y}} - \boldsymbol{\bar{\bar{x}}'b} \qquad \hat{\alpha}_i = a_i = (\bar{y}_i - \bar{\bar{y}}) - (\boldsymbol{\bar{x}_i} - \boldsymbol{\bar{\bar{x}}})'\boldsymbol{b}$$

$$\hat{\delta}_t = d_t = (\bar{y}_t - \bar{\bar{y}}) - (\boldsymbol{\bar{x}_t} - \boldsymbol{\bar{\bar{x}}})'\boldsymbol{b}$$

$$\bar{y}_t = \frac{1}{n}\sum_{i=1}^{n} y_{it} \qquad \bar{\bar{y}} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it}$$

$$\boldsymbol{\bar{x}_t} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_{it}} \qquad \boldsymbol{\bar{\bar{x}}} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} \boldsymbol{x_{it}}$$

- La matriz de covarianzas asintótica para $\boldsymbol{b}$ se puede estimar utilizando las sumas cuadradas y productos cruzados de $\boldsymbol{x_{it}^*}$, y s^2 se puede calcular con la siguiente fórmula:

$$s^2 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \boldsymbol{x_{it}'\beta} - m - a_i + d_t)^2}{nT - (n-1) - (T-1) - K - 1}$$

- Este modelo no se suele utilizar en la práctica porque el coste en términos de grados de libertad no suele ser justificado (s^2 es muy pequeño comparado con la formulación asimétrica) y se puede utilizar un modelo más general para modelar la evolución temporal del error

- Además, al incluir los efectos temporales, se puede ver que la estimación por primeras diferencias no permitirá eliminar estas para una estimación de mínimos cuadrados ordinarios

- Si los regresores no están correlacionados en el sentido estricto con los efectos individuales, entonces los efectos de la heterogeneidad provienen del término de error y lo más apropiado es modelar los términos constantes específicos como aleatorios para cada unidad, de modo que se puede utilizar el modelo de efectos aleatorios

    - El modelo de efectos fijos permite que los efectos de los individuos no observados estén correlacionados con las variables incluidas, modelando las diferencias entre unidades como desplazamientos del la

función de regresión (por las variables binarias) con el término α_i que proviene de c_i (la heterogeneidad no observada)

- Este modelo puede entenderse como uno que solo aplica para las unidades transversales en el estudio, pero no para unidades observacionales fuera de él (los efectos fijos son para los individuos de la muestra, no para individuos fuera de ella)

- El enfoque de parámetros aleatorios es más útil si se considera que la muestra transversal se ha sacado de una población grande (dado que el efecto fijo específico de los individuos sería como aleatorio debido al tamaño de la población), dado que los efectos específicos de cada unidad se distribuirán de manera equivalente para todas las unidades (haciendo que los efectos individuales de la heterogeneidad no observable sean globales y no se puedan diferenciar entre individuos como antes)

- Haciendo esto, es posible reducir el número de parámetros a estimar, aunque el coste es la posibilidad de tener estimador inconsistentes si las suposiciones son inapropiadas

- Si se reconsidera una reformulación del modelo visto anteriormente, pero se incluye la constante en los K regresores, un término constante α y un término aleatorio u_i, entonces la formulación del modelo es la siguiente:

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + E(\boldsymbol{z}_i'\boldsymbol{\alpha}|\boldsymbol{x}_{it}) + \boldsymbol{z}_i'\boldsymbol{\alpha} - E(\boldsymbol{z}_i'\boldsymbol{\alpha}|\boldsymbol{x}_{it}) + \varepsilon_{it}$$

$$\Rightarrow\ y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + E(c_i|\boldsymbol{x}_{it}) + c_i - E(c_i|\boldsymbol{x}_{it}) + \varepsilon_{it}$$

$$\Rightarrow\ y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \alpha + (u_i + \varepsilon_{it})$$

  - En este modelo, se asume que los efectos individuales específicos para cada unidad se distribuyen alrededor de una media común $\alpha \equiv E(\boldsymbol{z}_i'\boldsymbol{\alpha}|\boldsymbol{x}_{it})$ acorde a una distribución de probabilidad

  - El término constante α es la media de la heterogeneidad no observada (constante en el tiempo) y un término aleatorio $u_i \equiv \boldsymbol{z}_i'\boldsymbol{\alpha} - E(\boldsymbol{z}_i'\boldsymbol{\alpha}|\boldsymbol{x}_{it})$ es la heterogeneidad aleatoria específica o la variación del efecto individual específico para la observación i con respecto a su media (constante en el tiempo), de modo que se puede interpretar como la incertidumbre introducida por la heterogeneidad no observable y aleatoria de los individuos

  - Aunque se incluye la constante en los K regresores, la constante final del modelo será $\beta_0 + \alpha$, dado que no se pueden diferenciar

ambas a la hora de estimar (tal como pasa al estimar el modelo de efectos fijos con variables binarias, ya que el intercepto es $\beta_0 + \alpha_{-i}$ para aquella binaria que se omite)

- Si se asume exogeneidad estricta, entonces se obtienen las siguientes identidades:

$$E(\varepsilon_{it}|\boldsymbol{X}) = 0 \qquad E(u_{it}|\boldsymbol{X}) = 0$$

$$E(\varepsilon_{it}^2|\boldsymbol{X}) = \sigma_\varepsilon^2 \qquad E(u_i^2|\boldsymbol{X}) = \sigma_u^2$$

$$E(\varepsilon_{it}u_j|\boldsymbol{X}) = 0 \quad for\ all\ i,t\ and\ j$$

$$E(\varepsilon_{it}\varepsilon_{js}|\boldsymbol{X}) = 0 \quad if\ \ t \neq s\ or\ i \neq j$$

$$E(u_u u_j|\boldsymbol{X}) = 0 \quad if\ \ i \neq j$$

- Igual que antes, es útil reformular el modelo en bloques de T a través de vectores para el grupo i

$$\boldsymbol{y_i} = \boldsymbol{X_i'\beta} + \alpha + u_i\boldsymbol{i} + \boldsymbol{\varepsilon_t}$$

- Para estas T observaciones se define un término de error compuesto $\eta_{it} \equiv \varepsilon_{it} + u_i$, cuyo vector para las T observaciones es $\boldsymbol{\eta_i'} = (\eta_{i1}, \eta_{i2}, \dots, \eta_{iT})$ y hace que haya dos fuentes de perturbaciones

$$\boldsymbol{y_i} = \boldsymbol{X_i'\beta} + \alpha + \boldsymbol{\eta_i}$$

- A partir de esta forma de $\boldsymbol{\eta_i}$, el modelo se puede reformular, y se llama modelo de componentes del error. En este modelo, se cumplen las siguientes igualdades:

$$E(\eta_{it}\eta_{is}|\boldsymbol{X}) = \sigma_\varepsilon^2 + \sigma_u^2$$

$$E(\eta_{it}\eta_{is}|\boldsymbol{X}) = \sigma_u^2 \quad for\ t \neq s$$

$$E(\eta_{it}\eta_{js}|\boldsymbol{X}) = 0 \quad for\ all\ t\ and\ s\ if\ i \neq j$$

- Para estas T observaciones para la unidad observacional i, siendo $\boldsymbol{\Sigma} \equiv E(\boldsymbol{\eta_i}\boldsymbol{\eta_i'}|\boldsymbol{X})$ la matriz de covarianzas, $\boldsymbol{i_T}$ un vector $T \times 1$ de unos e $\boldsymbol{I_T}$ una matriz identidad $T \times T$, esta se puede expresar de la siguiente manera:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ & \cdots & & & \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_\varepsilon^2 + \sigma_u^2 \end{bmatrix} = \sigma_\varepsilon^2 \mathbf{I}_T + \sigma_u^2 \mathbf{i}_T \mathbf{i}_T'$$

- Debido a que las observaciones i y j son independientes, la matriz de covarianzas de las perturbaciones para las nT observaciones es la siguiente:

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ & & & \vdots & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{bmatrix} = \mathbf{I}_n \otimes \boldsymbol{\Sigma}$$

- El modelo anterior es un modelo de regresión generalizado como los vistos anteriormente, por lo que las implicaciones anteriormente vistas del modelo de generalizado también aplican

  - Las perturbaciones están autocorrelacionadas (en el tiempo) dentro de un mismo grupo, pero no entre grupos. Esta autocorrelación de y_{it} entre periodos equivale a la fracción de la varianza de y_{it} explicada por los efectos aleatorios

$$Corr(y_{it}, y_{is}) = \frac{E(\eta_{it}\eta_{is}|\boldsymbol{X})}{\sqrt{E(\eta_{it}\eta_{is}|\boldsymbol{X})}\sqrt{E(\eta_{it}\eta_{is}|\boldsymbol{X})}} = \frac{\sigma_u^2}{\sigma_\varepsilon^2 + \sigma_u^2}$$

$$Fraction\ of\ variance\ of\ rand.effects = \frac{\sigma_u^2}{\sigma_\varepsilon^2 + \sigma_u^2}$$

  - En particular, los parámetros del modelo de efectos aleatorios se pueden estimar consistentemente mediante estimadores de mínimos cuadrados ordinarios, aunque no de manera eficiente. Además, la matriz de covarianzas se puede estimar asintóticamente y de manera robusta a través del estimador visto anteriormente para la del modelo agrupado

- No obstante, existen otros estimadores consistentes disponibles, tales como el estimador LSDV o el de medias grupales

  - Asumiendo que no hay regresores invariantes en el tiempo, se pueden considerar las desviaciones de la media con respecto al tiempo y el estimador LSDV sería un estimador consistente de $\boldsymbol{\beta}$. No obstante, este estimador es ineficiente

$$y_{it} - \bar{y}_i = (\boldsymbol{x_{it}} - \bar{\boldsymbol{x}}_{\boldsymbol{i}})'\boldsymbol{\beta} + \alpha + \varepsilon_{it} - \bar{\varepsilon}_i$$

- Considerando la media grupal, se puede estimar consistentemente $\boldsymbol{\beta}$ a través del estimador de medias grupales, aunque este estima de manera ineficiente

$$\bar{y}_i = \bar{\boldsymbol{x}}_{\boldsymbol{i}}'\boldsymbol{\beta} + \alpha + u_i + \bar{\varepsilon}_i$$

- Aunque no sean los estimadores más eficientes, como la mayoría de modelos generalizados se estiman en dos partes, estos estimadores pueden servir como un estimador robusto de mínimos cuadrados para los parámetros de la varianza en el modelo. Los siguientes estimadores permiten obtener estimadores consistentes de funciones de varianzas:

$$\boldsymbol{e}_{\boldsymbol{pooled}}'\boldsymbol{e}_{\boldsymbol{pooled}}/nT \xrightarrow{P} \sigma_u^2 + \sigma_\varepsilon^2$$

$$\boldsymbol{e}_{\boldsymbol{LSDV}}'\boldsymbol{e}_{\boldsymbol{LSDV}}/(n(T-1)-K) \xrightarrow{P} \sigma_\varepsilon^2$$

$$\boldsymbol{e}_{\boldsymbol{FD}}'\boldsymbol{e}_{\boldsymbol{FD}}/n(T-1) \xrightarrow{P} 2\sigma_\varepsilon^2$$

$$\boldsymbol{e}_{\boldsymbol{between}}'\boldsymbol{e}_{\boldsymbol{between}}/nT \xrightarrow{P} \sigma_u^2 + \sigma_\varepsilon^2/T$$

- Diferentes pares de estos estimadores (y otros candidatos) pueden proporcionar un estimador por método de momentos de dos ecuaciones para $(\sigma_\varepsilon^2, \sigma_u^2)$, por lo que se puede desarrollar un estimador de mínimos cuadrados generalizados eficiente

- Para obtener el estimador de mínimos cuadrados generalizados de la pendiente en este contexto, es necesario transformar los datos y aplicar mínimos cuadrados ordinarios

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y} = \left(\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_i\right)$$

  - Se requiere que $\boldsymbol{\Omega}^{-1/2} = (\boldsymbol{I}_n \otimes \boldsymbol{\Sigma})^{-1/2} = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}^{-1/2}$ (la matriz formada por el producto de cada uno de los componentes de $\boldsymbol{I}_n$ por la matriz entera $\boldsymbol{\Sigma}^{-1/2}$), de modo que solo es necesario encontrar $\boldsymbol{\Sigma}^{-1/2}$ y esta se define de la siguiente manera:

$$\boldsymbol{\Sigma}^{-1/2} = \left[\boldsymbol{I}_T - \frac{\theta_i}{T}\boldsymbol{i}_{\boldsymbol{T}}\boldsymbol{i}_{\boldsymbol{T}}'\right] \quad where \;\; \theta_i = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_u^2}}$$

  - La transformación de $\boldsymbol{y}_{\boldsymbol{i}}$ y $\boldsymbol{X}_{\boldsymbol{i}}$ para el método de mínimos cuadrados generalizados es, por tanto, la siguiente:

$$\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y_i} = \frac{1}{\sigma_\varepsilon}\begin{bmatrix} y_{i1} - \theta\overline{y}_i \\ y_{i2} - \theta\overline{y}_i \\ \cdots \\ y_{iT} - \theta\overline{y}_i \end{bmatrix} \qquad \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X_i} = \frac{1}{\sigma_\varepsilon}\begin{bmatrix} \boldsymbol{x}_{i1} - \theta\overline{\boldsymbol{x}}_{\boldsymbol{i}} \\ \boldsymbol{x}_{i2} - \theta\overline{\boldsymbol{x}}_{\boldsymbol{i}} \\ \cdots \\ \boldsymbol{x}_{iT} - \theta\overline{\boldsymbol{x}}_{\boldsymbol{i}} \end{bmatrix}$$

- En el caso en el que $\theta = 1$, el procedimiento es equivalente al de estimar a través del estimador LSDV (el estimado para un modelo de efectos fijos), mientras que si $\theta = 0$, se obtiene un estimador de mínimos cuadrados ordinarios (como en el modelo agrupado). Esto ocurre porque θ se puede entender como la raíz de la fracción de la desviación estándar explicada por el término u_i (para expresarse en las mismas unidades de y_{it})

$$\theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_u^2}} = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2}} = \sqrt{\frac{\sigma_u^2}{\sigma_\varepsilon^2 + T\sigma_u^2}}$$

- Se puede demostrar que el estimador de mínimos cuadrados generalizados es una media ponderada matricial de los estimadores dentro de grupos y entre grupos:

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{F}}^{\boldsymbol{within}}\boldsymbol{b}^{\boldsymbol{within}} + \left(\boldsymbol{I} - \widehat{\boldsymbol{F}}^{\boldsymbol{within}}\right)\boldsymbol{b}^{\boldsymbol{between}}$$

$$where \quad \widehat{\boldsymbol{F}}^{\boldsymbol{within}} = \left[\boldsymbol{S}_{\boldsymbol{xx}}^{\boldsymbol{within}} + \lambda\boldsymbol{S}_{\boldsymbol{xx}}^{\boldsymbol{between}}\right]^{-1}\boldsymbol{S}_{\boldsymbol{xx}}^{\boldsymbol{within}} \quad \&$$

$$\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2} = (1-\theta)^2$$

- Los modelos no balanceados complican el modelo de efectos aleatorios porque la matriz $\boldsymbol{\Omega} \neq \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}$ porque los bloques diagonales de $\boldsymbol{\Omega}$ son de diferentes tamaños. No obstante, la estimación debe verse inafectada dado que la fuente de heteroscedasticidad proviene de las diferencias de tamaños entre grupos

$$\boldsymbol{\Sigma}^{-1/2} = \left[\boldsymbol{I}_{T_i} - \frac{\theta_i}{T_i}\boldsymbol{i}_{T_i}\boldsymbol{i}'_{T_i}\right] \quad where \quad \theta_i = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T_i\sigma_u^2}}$$

- Mientras λ difiera de 1, la ineficiencia de los mínimos cuadrados ordinarios y del estimador dentro de los grupos usado para el modelo de efectos fijos provendrá de una ponderación ineficiente de los dos estimadores

  - Comparado con el estimador de mínimos cuadrados generalizados, el estimador MCO pone demasiado peso en las unidades de variación entre grupos (incluye toda la variación en

$\boldsymbol{X}$ y no tiene en cuenta que parte de la variación entre grupos proviene de la variación en u_i entre unidades)

- En cuanto al modelo de efectos fijos, se puede ver como, de atribuir la variabilidad a más de una fuente, el estimador dentro de los grupos no es eficiente porque no tiene en cuenta σ_ε^2 y atribuye todo el peso a $\boldsymbol{b^{within}}$

- Por lo tanto, cuanto mayor sea σ_u^2, más cerca se estará del estimador usado en el modelo de efectos fijos, y mientras menor sea σ_u^2, más cerca se estará del estimador de mínimos cuadrados ordinarios

- El enfoque de variables binarias es costoso en términos de grados de libertad perdidos, mientras que la suposición de no correlación entre efectos individuales y regresores no está muy justificada. Por lo tanto, para poder saber si usar el modelo de efectos fijos o de efectos aleatorios, se puede utilizar el contraste de especificación de Hausman

  - Este contrasta la ortogonalidad de los efectos comunes y de los regresores, basándose en la idea de que, bajo la hipótesis nula de no correlación, tanto el estimador LSDV como el FGLS deben ser consistentes, pero el LSDV debe ser ineficiente (se prefiere un modelo de efectos aleatorios). La hipótesis alternativa es que el estimador LSDV es consistente pero el estimador FGLS no lo es (se prefiere el modelo de efectos fijos)

  - Por lo tanto, bajo la hipótesis nula, las dos estimaciones no deberían diferir sistemáticamente, y se puede hacer un contraste para esa diferencia

  - Otro ingrediente esencial para el contraste es la matriz de varianzas y covarianzas del vector $(\boldsymbol{b_{FE}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{RE}})$. La covarianza entre un estimador eficiente y su diferencia de un estimador ineficiente debe ser nula, lo cual implica que la covarianza entre el estimador eficiente e ineficiente es la varianza del ineficiente

$$Var\left(\boldsymbol{b_{FE}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{RE}}\right) = Var(\boldsymbol{b_{FE}}) + Var\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{RE}}\right) - 2Cov\left(\boldsymbol{b_{FE}}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{RE}}\right)$$

$$Cov\left[\left(\boldsymbol{b_{FE}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{RE}}\right), \widehat{\boldsymbol{\beta}}_{\boldsymbol{RE}}\right] = Cov\left(\boldsymbol{b_{FE}}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{RE}}\right) - Var\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{RE}}\right) = 0$$

  - Insertando este resultado en la ecuación, se obtiene la matriz de varianzas y covarianzas para el contraste de Wald. Para $\widehat{\boldsymbol{\Psi}}$ se utiliza la matriz de varianzas y covarianzas del estimador de las pendientes en el modelo LSDV y la matriz estimada en el modelo de efectos aleatorios (excluyendo el término constante)

$$Var\left(\boldsymbol{b_{FE}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{RE}}\right) = Var(\boldsymbol{b_{FE}}) - Var\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{RE}}\right) = \boldsymbol{\Psi}$$

$$W = \left(\boldsymbol{b_{FE}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{RE}}\right)' \widehat{\boldsymbol{\Psi}}^{-1} \left(\boldsymbol{b_{FE}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{RE}}\right) \sim \chi^2_{K-1}$$

- El contraste de Hausman es una herramienta útil para determinar la especificación preferida del modelo de efectos comunes. Un problema de este contraste, sin embargo, es que el estadístico puede ser negativo (no se garantiza que la diferencia sea definida positiva), por lo que se puede considerar la siguiente forma del contraste, la cual asegura una matriz definida no positiva y $\boldsymbol{b_{means}}$ es el estimador de medias entre grupos:

$$H' = (\boldsymbol{b_{FE}} - \boldsymbol{b_{means}})'[Asy.Var(\boldsymbol{b})]^{-1}(\boldsymbol{b_{FE}} - \boldsymbol{b_{means}})$$

$$where \;\; Asy.Var(\boldsymbol{b}) = Asy.Var(\boldsymbol{b_{FE}}) + Asy.Var(\boldsymbol{b_{means}})$$

- Escoger entre el modelo de efectos fijos o aleatorios es un dilema, dado que ambas tienen aspectos negativos, tales como la proliferación de parámetros y los regresores invariantes en el tiempo para el modelo de efectos fijos, o la no correlación de la heterogeneidad no observada con los regresores del modelo de efectos aleatorios

  - Mundlak, en su investigación de 1978, suigirió una especificación para la esperanza de la heterogeneidad no observada, de modo que el modelo de efectos aleatorios queda de la siguiente manera:

$$E(c_i|\boldsymbol{X_i}) = \overline{\boldsymbol{x}}_{\boldsymbol{i}}'\boldsymbol{\gamma}$$

$$y_{it} = \boldsymbol{x}_{\boldsymbol{it}}'\boldsymbol{\beta} + c_i + \varepsilon_{it} = \boldsymbol{x}_{\boldsymbol{it}}'\boldsymbol{\beta} + \overline{\boldsymbol{x}}_{\boldsymbol{i}}'\boldsymbol{\gamma} + \varepsilon_{it} + \left(c_i - E(c_i|\boldsymbol{X_i})\right) =$$

$$= \boldsymbol{x}_{\boldsymbol{it}}'\boldsymbol{\beta} + \overline{\boldsymbol{x}}_{\boldsymbol{i}}'\boldsymbol{\gamma} + \varepsilon_{it} + u_i$$

  - Esta preserva la especificación del modelo de efectos aleatorios, pero no se puede lidiar con el problema de la correlación de los efectos y los regresores

  - El término $\overline{\boldsymbol{x}}_{\boldsymbol{i}}'\boldsymbol{\gamma}$ solo incluye variables que varían en el tiempo, y la diferencia entre el modelo de efectos fijos y el de efectos aleatorios es el vector diferente de cero $\boldsymbol{\gamma}$. De este modo, un test estadístico para $\boldsymbol{\gamma} = \boldsymbol{0}$ proporciona un enfoque alternativo para el test de Hausman

- El tratamiento de los datos de panel hasta ahora ha asumido que los parámetros del modelo son constantes, y los interceptos varían aleatoriamente

entre los grupos, de modo que la heterogeneidad se introduce a través de la variación del término constante

- La heterogeneidad de los parámetros entre individuos o grupos se puede modelar como una variación estocástica. Por lo tanto, se puede plantear el siguiente modelo:

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta_i} + \boldsymbol{\varepsilon_i} \qquad E(\boldsymbol{\varepsilon_i}|\boldsymbol{X_i}) = \mathbf{0} \qquad E(\boldsymbol{\varepsilon_i}\boldsymbol{\varepsilon_i'}|\boldsymbol{X_i}) = \sigma_\varepsilon^2 \boldsymbol{I}_T$$

$$\boldsymbol{\beta_i} = \boldsymbol{\beta} + \boldsymbol{u_i} \qquad E(\boldsymbol{u_i}|\boldsymbol{X_i}) = \mathbf{0} \qquad E(\boldsymbol{u_i}\boldsymbol{u_i'}|\boldsymbol{X_i}) = \boldsymbol{\Gamma}$$

  - Asumiendo que no hay autocorrelación o correlación transversal entre observaciones de las perturbaciones y que $T > K$ (se puede obtener la regresión para cada grupo), entonces $\boldsymbol{\beta_i}$ es la realización de un proceso aleatorio con media $\boldsymbol{\beta}$ y matriz de varianzas y covarianzas $\boldsymbol{\Gamma}$

  - Expandiendo los resultados en el modelo de regresión, se puede obtener el siguiente modelo general:

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + (\boldsymbol{\varepsilon_i} + \boldsymbol{X_i}\boldsymbol{u_i})$$

$$\boldsymbol{\Omega_{ii}} = E[(\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta})(\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta})'|\boldsymbol{X_i}] = \sigma_\varepsilon^2 \boldsymbol{I}_T + \boldsymbol{X_i}\boldsymbol{\Gamma}\boldsymbol{X_i'}$$

  - Para el sistema completo, la matriz de varianzas y covarianzas de las perturbaciones es diagonal en bloque, con un bloque diagonal $\boldsymbol{\Omega_{ii}}$ de tamaño $T \times T$

- Se puede escribir el estimador GLS como una media ponderada de matrices de los estimadores mínimos cuadrados ordinarios específicos para cada grupo

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y} = \sum_{i=1}^{n} \boldsymbol{W_i}\boldsymbol{b_i}$$

$$where \quad \boldsymbol{W_i} = \left[\sum_{i=1}^{n} (\boldsymbol{\Gamma} + \sigma_\varepsilon^2(\boldsymbol{X_i'}\boldsymbol{X_i})^{-1})^{-1}\right]^{-1} (\boldsymbol{\Gamma} + \sigma_\varepsilon^2(\boldsymbol{X_i'}\boldsymbol{X_i})^{-1})^{-1}$$

  - La implementación del modelo requiere un estimador para $\boldsymbol{\Gamma}$. Un enfoque es utilizar la varianza empírica del conjunto n de estimadores de mínimos cuadrados menos el valor medio de $s_i^2(\boldsymbol{X_i'}\boldsymbol{X_i})^{-1}$

$$\boldsymbol{G} = [1/(n-1)]\left(\sum_i \boldsymbol{b_i}\boldsymbol{b_i'} - n\overline{\boldsymbol{b}}\,\overline{\boldsymbol{b}}'\right) - (1/N)\sum_i \boldsymbol{V}_i$$

$$\bar{\boldsymbol{b}} = \frac{1}{n}\sum_i \boldsymbol{b_i} \qquad \boldsymbol{V}_i = s_i^2(\boldsymbol{X_i'X_i})^{-1}$$

- No obstante, la matriz puede no ser definida positiva, de modo que se tendría que sacar el segundo término

- Un contraste chi cuadrado para contrastar el modelo de coeficientes aleatorios ante la alternativa de la regresión clásica (sin aleatoriedad de los coeficientes) se puede basar en lo siguiente:

$$C = \sum_i (\boldsymbol{b_i} - \boldsymbol{b}_*)'\boldsymbol{V_i}(\boldsymbol{b_i} - \boldsymbol{b}_*) \sim \chi^2_{(n-1)K}$$

$$where \quad \boldsymbol{b}_* = \left(\sum_i \boldsymbol{V_i^{-1}}\right)^{-1} \sum_i \boldsymbol{V_i^{-1}}\boldsymbol{b_i}$$

- El mejor estimador lineal no sesgado para predictores individuales de los vectores de coeficientes de grupos específicos es una media ponderada de matrices del estimador GLS y del estimador de mínimos cuadrados

$$\widehat{\boldsymbol{\beta}}_i = \boldsymbol{Q_i}\widehat{\boldsymbol{\beta}} + (\boldsymbol{I} - \boldsymbol{Q_i})\boldsymbol{b_i}$$

$$where \quad \boldsymbol{Q_i} = [(1/s_i^2)\boldsymbol{X_i'X_i} + \boldsymbol{G}^{-1}]^{-1}\boldsymbol{G}^{-1}$$

  - Muchos investigadores utilizan un enfoque de dos pasos para estimar modelos de dos niveles. En una forma común de la aplicación, se utiliza un conjunto de datos de panel para poder estimar el siguiente modelo:

$$\boldsymbol{y_{it}} = \boldsymbol{x_{it}'\beta_i} + \boldsymbol{\varepsilon_{it}} \quad for \;\; i = 1, \dots, n \;\; and \;\; t = 1, \dots, T$$

$$\beta_{i,k} = \boldsymbol{z_i'\alpha_k} + \boldsymbol{u_{i,k}} \quad for \;\; i = 1, \dots, n$$

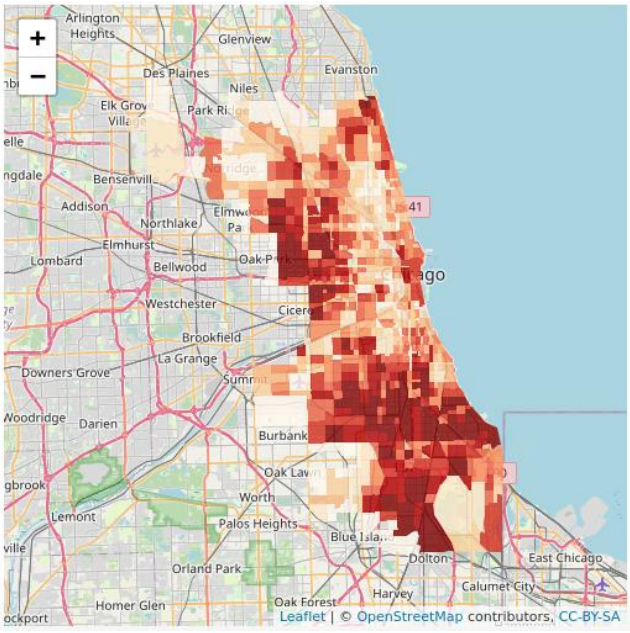
    - Asumiendo que el panel es lo suficientemente grande, la primera ecuación se estima n veces (una para cada individuo) y entonces el coeficiente estimado de x_{itk} en cada regresión forma una observación para la regresión en el segundo paso

- Los efectos de la clusterización mencionados anteriormente se justifican por la esperanza de que los efectos en localizaciones vecinas se comparten entre estas, creando una especie de correlación entre el espacio (más que en el tiempo) la cual se denomina autocorrelación espacial

- La autocorrelación espacial mide el grado en el que un fenómeno de interés está correlacionado a si mismo en el espacio

    - En otras palabras, valores similares aparecen más juntos uno al otro (o clusterizan) en el espacio (autocorrelación espacial positiva) o los valores vecinos son diferentes (autocorrelación espacial negativa). Una autocorrelación espacial nula indica que el patrón espacial es aleatorio (el espacio no afecta)

    - Para poder tener una idea visual sobre si es necesario considerar la autocorrelación, se puede utilizar un gráfico (mapa) que indique la densidad de los valores de la variable dependiente para cada localización. De este modo, si se puede identificar algún patrón según la cercanía, podría ser necesario considerar la autocorrelación espacial en los datos



    - Es posible expresar la existencia de autocorrelación espacial para dos observaciones de una variable aleatoria y_i y y_j en las localizaciones i y j

$$Cov(y_i, y_j) \neq 0 \quad for \;\; i \neq j$$

- Un modelo de autocorrelación espacial toma la misma forma familiar de las regresiones de datos de panel, pero se define el error de manera diferente para inducir autocorrelación espacial. Este tipo de modelos que introducen la autocorrelación a través del error se denominan modelos de errores espaciales (SEM)

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} + u_i \quad for \;\; i = 1,2,\dots,n \;\; and \;\; t = 1,2,\dots,T$$

- En esta representación, el subíndice i representa los individuos o grupos como antes, pero en este caso se pueden interpretar también como localizaciones. El término u_i sigue representando los efectos individuales o de grupos, pero ε_{it} permite que se tenga una estructura de autocorrelación espacial si se define de la siguiente manera:

$$\varepsilon_{it} = \lambda \sum_{j=1}^{n} W_{ij}\varepsilon_{jt} + \nu_t \quad for \quad i = 1,2, \dots, n$$

- El escalar λ es el coeficiente de autocorrelación espacial, y los elementos W_{ij} son ponderaciones espaciales o contiguas que se conocen (por suposición). Los elementos que aparecen en la suma de arriba serán una fila del peso espacial o matriz de contigüidad $\boldsymbol{W}$ de tamaño $n \times n$, de modo que para las n unidades se tiene la siguiente igualdad:

$$\boldsymbol{\varepsilon_t} = \lambda \boldsymbol{W}\boldsymbol{\varepsilon_t} + \boldsymbol{\nu_t} = \lambda \boldsymbol{W}\boldsymbol{\varepsilon_t} + \nu_t \boldsymbol{i}$$

- Asumiendo que $|\lambda| < 1$ y que los elementos de la matriz $\boldsymbol{W}$ son tales que $(\boldsymbol{I}_n - \lambda \boldsymbol{W})$ sea una matriz invertible, es posible reescribir $\boldsymbol{\varepsilon_t}$ como $\boldsymbol{\varepsilon_t} = (\boldsymbol{I}_n - \lambda \boldsymbol{W})^{-1}\boldsymbol{\nu_t}$ y expresar el modelo de la siguiente manera:

$$\boldsymbol{y}_t = \boldsymbol{X_t}\boldsymbol{\beta} + (\boldsymbol{I}_n - \lambda \boldsymbol{W})^{-1}\boldsymbol{\nu_t} + \boldsymbol{u}$$

- Adicionalmente, se asume que u_i y v_i tienen medias nulas y varianzas constantes σ_u^2 y σ_v^2 y que son independientes entre individuos y entre ellos. Por lo tanto, el modelo de regresión generalizado que aplica a las n observaciones en un momento t será el siguiente:

$$E(\boldsymbol{y}_t|\boldsymbol{X_t}) = \boldsymbol{X_t}\boldsymbol{\beta}$$

$$Var(\boldsymbol{y}_t|\boldsymbol{X_t}) = (\boldsymbol{I}_n - \lambda \boldsymbol{W})^{-1}(\sigma_v^2 \boldsymbol{i}\boldsymbol{i}')(\boldsymbol{I}_n - \lambda \boldsymbol{W})^{-1} + \sigma_u^2 \boldsymbol{I_n}$$

- La estimación y la interpretación se puede hacer como en otros modelos de datos de panel, aunque no hay un estimador natural para λ (a veces se asume normalidad y se usan métodos de máxima verosimilitud)

- Otro tipo de modelo muy utilizado es uno que funciona como un modelo autorregresivo para series temporales, el cual se denomina modelos espaciales autorregresivos (SAR). Este tipo de modelos fueron propuestos por Anselin en 1988 como extensiones del modelo espacial a regresiones dinámicas, y hay varias especificaciones

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it} + u_i \quad for \quad i = 1,2,\dots,n \quad and \quad t = 1,2,\dots,T$$

- Un modelo recursivo de espacio puro o *pure space-recursive model* especifica que la autocorralación pertenece a los vecinos en el periodo anterior

$$y_{it} = \gamma[\boldsymbol{W}\boldsymbol{y}_{t-1}]_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}$$

$$for \quad i = 1,2,\dots,n \quad and \quad t = 1,2,\dots,T$$

- Un modelo recursivo de espacio-tiempo o *time-space recursive model* especifica que la dependencia es puramente autorregresiva respecto a los vecinos en el periodo pasado

$$y_{it} = \rho y_{i,t-1} + \gamma[\boldsymbol{W}\boldsymbol{y}_{t-1}]_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}$$

$$for \quad i = 1,2,\dots,n \quad and \quad t = 1,2,\dots,T$$

- Un modelo simultáneo de espacio-tiempo o *time-space simultaneous model* especifica que la dependencia espacial es con respecto a los vecinos en el periodo actual

$$y_{it} = \rho y_{i,t-1} + \lambda[\boldsymbol{W}\boldsymbol{y}_{t}]_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}$$

$$for \quad i = 1,2,\dots,n \quad and \quad t = 1,2,\dots,T$$

- Finalmente, un modelo dinámico de espacio-tiempo o *time-space dynamic model* especifica que la autorregresión depende de los vecinos en el periodo actual y en el anterior

$$y_{it} = \rho y_{i,t-1} + \lambda[\boldsymbol{W}\boldsymbol{y}_{t}]_i + \gamma[\boldsymbol{W}\boldsymbol{y}_{t-1}]_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}$$

$$for \quad i = 1,2,\dots,n \quad and \quad t = 1,2,\dots,T$$

- En este caso, serán necesarios métodos de estimación para modelos de datos de panel y la interpretación de los coeficientes en la regresión variarán según el término

- El coeficiente γ es el coeficiente autorregresivo espacial para el periodo anterior, mientras que λ es el mismo coeficiente, pero para el periodo actual (son las correlaciones). El signo del coeficiente indica la dirección de la autocorrelación espacial. El coeficiente ρ se puede interpretar como el efecto del valor pasado en el presente (como en los modelos autorregresivos)

- Los modelos presentados sirven para modelar datos de panel, pero también tienen una contraparte para datos de sección cruzada:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \;\; where \;\; \varepsilon_i = \; \lambda \sum_{j=1}^{n} W_{ij}\varepsilon_j + \nu_i \;\; for \;\; i = 1,2,\dots,n \;\; (SEM)$$

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \; \lambda \sum_{j=1}^{n} W_{ij} y_j + \varepsilon_i \;\; for \;\; i = 1,2,\dots,n \;\; (SAR)$$

- Con notación matricial, los modelos se pueden expresar equivalentemente:

$$\boldsymbol{y} = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\varepsilon} \;\; where \;\; \boldsymbol{\varepsilon} = \; \lambda \boldsymbol{W}\boldsymbol{\varepsilon} + \boldsymbol{v} = \lambda \boldsymbol{W}\boldsymbol{\varepsilon} + \nu_i \boldsymbol{i} \;\; (SEM)$$

$$\boldsymbol{y} = \boldsymbol{x}_i'\boldsymbol{\beta} + \lambda \boldsymbol{W}\boldsymbol{y} + \boldsymbol{\varepsilon} = \boldsymbol{x}_i'\boldsymbol{\beta} + \lambda \boldsymbol{W}\boldsymbol{y} + \varepsilon_i \boldsymbol{i} \;\; (SAR)$$

- Para el modelo SAR presentado para datos cruzados, el término $\boldsymbol{Wy}$ induce a una correlación no nula con el término de error, ya que los términos siempre están correlacionados independientemente de la estructura de los errores. Esto implica que los estimadores MCO en el modelo no espacial estarán sesgados y serán inconsistentes

- Si el modelo SEM es real, entonces los estimadores MCO para el modelo no espacial no están sesgados, pero son ineficientes, dado que la correlación proviene de $\boldsymbol{\varepsilon}$

- En este caso, λ es el coeficiente de autocorrelación espacial en el modelo SAR y su signo es la dirección de la autocorrelación, mientras que en el modelo SEM la interpretación sigue siendo la misma que para un modelo común de datos cruzados

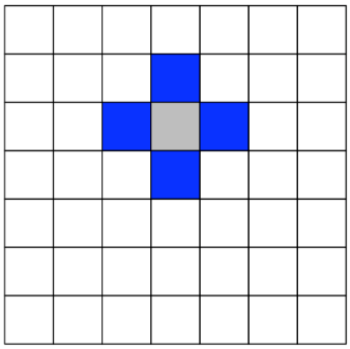
- No se consideran las otras especificaciones para los modelos SAR debido a que todas se diferencian al incluir diferentes momentos, y en datos cruzados el tiempo no se tiene en cuenta

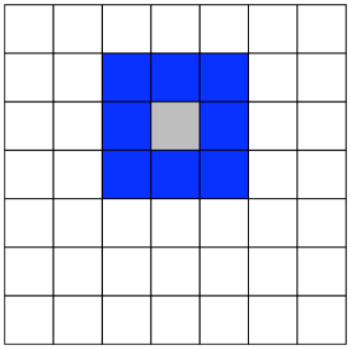
- La estructura del modelo proviene de la matriz simétrica de ponderaciones $\boldsymbol{W}$. Esta matriz puede representarse típicamente de varias maneras:

  - La primera manera consiste en que W_{ij} sea 1 cuando un par i, j son vecinos, mientras que será cero si no lo son para ese par. De este modo, en verdad se impone una restricción sobre cada punto, la cual es la definición de un conjunto de vecindad (en donde, por convención, no se incluyen los pares (i, i))

$$W_{ij} = \begin{cases} 1 & if \;\; (i,j) \in N(i,j) \\ 0 & otherwise \end{cases} \quad for \;\; \forall (i,j)$$

- La especificación del conjunto de vecindad es un poco arbitraria y hay bastantes maneras de poder hacerlo. Dos criterios muy utilizados son el de la torre (dos unidades son vecinas si comparten un costado de un cuadrado en el espacio) y el de la reina (dos unidades son vecinas si comparten un costado o una esquina)





- Otro criterio usado es el de clasificar dos observaciones o localizaciones como vecinas si están a una cierta distancia, de modo que se cumpla la siguiente condición:

$$(i,j) \in N(i,j) \;\; if \;\; |i-j| < d_{max}$$

- Finalmente, se puede hacer que W_{ij} sea una distancia en el espacio, de modo que W_{ij} decrece cuanto más aumente $|i-j|$ (la ponderación de la autocorrelación baja cuanto más aumenta la lejanía entre localizaciones). En este enfoque, las filas de cada matriz sumarán 1 (en otras palabras, $\sum_i W_{ij} = 1$)

- Un paso natural es el de utilizar un contraste hipótesis para los efectos espaciales. El procedimiento estándar para una regresión de datos de sección cruzada es el estadístico I de Moran, que se calcularía para cada conjunto de residuos $\boldsymbol{e_t}$

$$I_t = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(e_{it}-\bar{e}_t)(e_{jt}-\bar{e}_t)}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\right)\sum_{i=1}^{n}(e_{it}-\bar{e}_t)^2} = \frac{\boldsymbol{e_t' W e_t}}{\boldsymbol{e_t' e_t}}$$

- Como se puede ver, este estadístico es para conjuntos de cada $t = 1,2,\dots,T$, de modo que se puede obtener el mismo estadístico I para datos cruzados

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(e_i-\bar{e})(e_j-\bar{e})}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\right)\sum_{i=1}^{n}(e_i-\bar{e})^2} = \frac{\boldsymbol{e' W e}}{\boldsymbol{e' e}}$$

- El estadístico I de Moran es una medida de autocorrelación espacial, y este puede tomar valores entre -1 (dispersión perfecta) y 1 (correlación perfecta)

- Para un panel de T conjuntos de observaciones independientes, el estadístico $\bar{I} = \frac{1}{T}\sum_{t=1}^{T} I_t$ usaría todo el conjunto de información. Una aproximación asintótica para la varianza de este estadístico najo la hipótesis nula:

$$V^2 = \frac{1}{T}\frac{n^2\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}^2 + 3\left(\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}\right)^2 - n\sum_{i=1}^{n}\left(\sum_{j=1}^{n}W_{ij}\right)^2}{(n^2-1)\left(\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}\right)^2}$$

- El estadístico $\bar{I}/V$ se distribuirá como una distribución normal bajo la hipótesis nula y se puede utilizar un estadístico de contraste. La hipótesis nula que el estadístico $\bar{I}$ es nulo (lo cual quiere decir que la autocorrelación espacial es nula) mientras que tiene como hipótesis alternativa un valor diferente (habiendo así correlación espacial)

- Baltagi, Song and Koh, en su investigación de 2003, identificaron una variedad de contrastes de multiplicadores de Lagrange bajo la suposición de normalidad

  - Para datos cruzados y autocorrelación espacial en el error (para un modelo SEM), se puede utilizar el siguiente contraste, cuya hipótesis nula es que no hay un término $\boldsymbol{We}$ en el modelo real para y y la alternativa es que si hay (como en un SEM):

$$LM(1) = \frac{(\boldsymbol{e'We}/s^2)^2}{tr(\boldsymbol{W'W} + \boldsymbol{W}^2)} \quad where\ \ s^2 = \boldsymbol{e'e}/n$$

  - Para datos cruzados y para variables dependientes retrasadas (*lagged dependent variables*), se puede utilizar el siguiente contraste, cuya hipótesis nula es que no hay un término $\boldsymbol{Wy}$ en el modelo real para y y la alternativa es que si hay (como en un SAR):

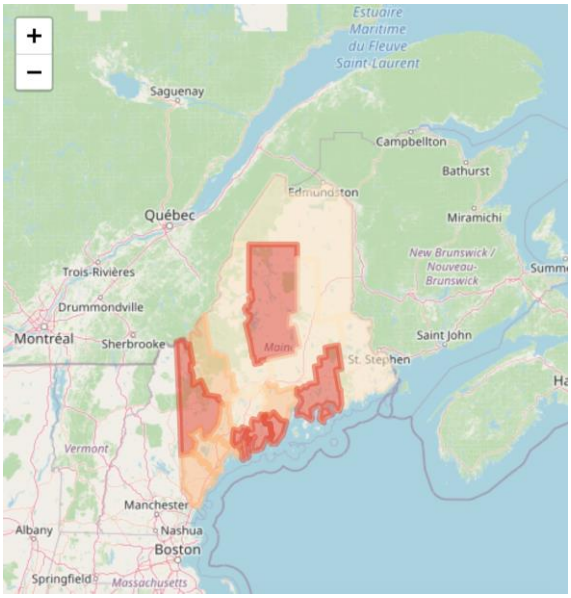
$$LM(2) = \frac{(\boldsymbol{e'Wy}/s^2)^2}{\boldsymbol{b'X'WMWXb}/s^2 + tr(\boldsymbol{W'W} + \boldsymbol{W}^2)}$$

$$where\ \ \boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$$

  - En este caso, rechazar la hipótesis nula quiere decir que hay evidencia sobre autocorrelación espacial (la hipótesis alternativa), mientras que no rechazar la hipótesis nula quiere

decir que la autocorrelación sería aleatoria (no habría autocorrelación espacial)

- Una manera visual de poder tener una idea del ajuste del modelo a los datos es a través de hacer un gráfico con la densidad de los residuos (como se hizo antes con la variable dependiente)



  - Si se detecta que hay algún patrón de clusterización para zonas con residuos significativos, entonces se puede interpretar como una señal de que el ajuste del modelo no es muy bueno porque falta algún factor que explique este patrón (autocorrelación espacial)

Los resultados binarios y la elección discreta

- El análisis de la elección individual que es foco de la microeconometría se basa en hacer modelar resultados discretos. De este modo, la variable dependiente no es una medida cuantitativa de un resultado económico, sino un indicador de que algo ha ocurrido o no

  - Por lo tanto, los modelos de regresión vistos hasta ahora no son apropiados, de modo que se tiene que pasar al modelaje de probabilidades y de conteo de ocurrencias

    - Los modelos estudiados en esta área son naturalmente no lineales

    - Estos se basan en un modelo de utilidad aleatoria para las elecciones observadas, de modo que el decisor se enfrenta a una situación o conjunto de alternativas y revela algo de sus preferencias subyacentes por la elección que hace. Estas

elecciones hechas se ven afectadas por influencias observables y por características inobservables del decisor

- Los modelos de elección discreta más comunes son el de elección binaria, el de elección multinomial, el de elección ordenada y el de conteo de eventos

  - En la elección binaria, los individuos se enfrentan a dos decisiones y se toma la decisión que proporciona la mayor utilidad. En este contexto, el valor de las variables binarias solo son indicadores numéricos para la elección

  - En la elección multinomial, los individuos se enfrentan a más de dos decisiones y se toma la decisión que proporciona la mayor utilidad. Esta es una variación pequeña de los modelos de elección binaria, pero modelos más elaborados permiten una mayor especificación de las preferencias de los individuos. Además, los valores de la variable dependiente son solo indicadores numéricos para la elección

  - En la elección ordenada, los individuos revelan la fuerza de sus preferencias con respecto a un único resultado. En este contexto, los valores de la variable tienen significado, dado que tienen un orden que revelan la fuerza de las preferencias, y la variable se considera una medida cuantitativa

  - En los modelos de conteo de eventos, el resultado observado es un conteo del número de ocurrencias de un evento. En muchos casos, estos son similar que, en los otros tres contextos, dado que la variable dependiente mide la elección individual, pero en otros casos, este conteo puede pertenecer a un proceso natural. Por lo tanto, hay contextos en los que se hace un modelaje de regresión familia, pero en los más generales se adapta el modelaje para acomodar una variable dependiente discreta y no negativa (a través de probabilidades)

- Con tal de estudia el comportamiento individual, se construyen modelos que vinculan la decisión o resultado a una serie de factores. El enfoque es analizar cada un en un marco general de modelos de probabilidad:

$$P(event\ j\ occurs|\boldsymbol{x_i}) = P(Y_i = j|\boldsymbol{x_i}) = F(relevant\ effects, parameters, \boldsymbol{x_i})$$

  - El estudio de la elección cualitativa se enfoca en una especificación y estimación apropiada, y utiliza modelos para las probabilidades de eventos, en donde los eventos son las elecciones de los individuos entre dos o más alternativas la mayoría de veces

- Por lo tanto, se denota por $P(Y_i = 1|\boldsymbol{x_i})$ la probabilidad de que el evento de interés ocurra dado $\boldsymbol{x_i}$ y $P(Y_i = 0|\boldsymbol{x_i})$ es la probabilidad de que el evento no ocurra

- Una interpretación de los datos sobre la elección individual es proporcionada por el modelo de utilidad aleatorio. Si U_a y U_b representan la utilidad de un individuo entre dos elecciones diferentes, la elección observada revela cuál de las dos opciones proporciona mayor utilidad, pero no la magnitud de la utilidad subyacente

  - Por lo tanto, el indicador observado es igual a 1 si $U_a > U_b$ y 0 si $U_a < U_b$, entonces $Y = \mathbf{1}(U > 0)$ si $U = U_a - U_b$ (lo mismo que el caso de censura visto anteriormente). Una formulación común es el modelo lineal de utilidad aleatoria:

$$U_a = \boldsymbol{w'\beta_a} + \boldsymbol{z'_a\gamma_a} + \varepsilon_a \qquad U_b = \boldsymbol{w'\beta_b} + \boldsymbol{z'_b\gamma_b} + \varepsilon_b$$

  - El vector observable de características del individuo se denota $\boldsymbol{w}$ (puede incluir ingresos, género, etc.). Los vectores $\boldsymbol{z_a}$ y $\boldsymbol{z_b}$ denota atributos de las dos elecciones (que pueden ser especificas de la elección), los términos aleatorios ε_a y ε_b representan los elementos estocásticos que son específicos y solo lo saben los individuos, pero no el observador

  - La completitud del modelo para la determinación de resultados observados es la revelación del *ranking* de las preferencias por la elección del individuo. Entonces, si se denota por $Y = 1$ la elección de consumo de la alternativa a, se infiere que $U_a > U_b$ y como el resultado depende de elementos aleatorios, entonces se obtiene el siguiente modelo:

$$P(Y = 1|\boldsymbol{w}, \boldsymbol{z}_a, \boldsymbol{z}_b) = P(U_a > U_b) =$$

$$= P[(\boldsymbol{w'\beta_a} + \boldsymbol{z'_a\gamma_a} + \varepsilon_a) - (\boldsymbol{w'\beta_b} + \boldsymbol{z'_b\gamma_b} + \varepsilon_b) > 0|\boldsymbol{w}, \boldsymbol{z}_a, \boldsymbol{z}_b] =$$

$$= P[\boldsymbol{w'}(\boldsymbol{\beta_a} - \boldsymbol{\beta_b}) - (\boldsymbol{z'_a\gamma_a} - \boldsymbol{z'_b\gamma_b}) + (\varepsilon_a + \varepsilon_b) > 0|\boldsymbol{w}, \boldsymbol{z}_a, \boldsymbol{z}_b] =$$

$$= P(\boldsymbol{x'\beta} + \varepsilon > 0|\boldsymbol{x})$$

  - En este caso $\boldsymbol{x'\beta}$ incluye todos los términos observables de la diferencia de las dos funciones de utilidad y ε denota la diferencia entre dos elementos aleatorios

- Los modelos de variable dependiente discreta normalmente se denominan modelos de función de índice. El resultado de una elección discreta es un reflejo de una regresión subyacente, modelando la diferencia entre la utilidad de las opciones con una regresión de la variable no observada y^*

$$y^* = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon$$

- Este resultado es la utilidad neta mostrada anteriormente, y se asume que ε tiene media nula (de modo que en $\boldsymbol{x}$ tiene que haber un término constante) y normalmente tiene una distribución logístico u otra distribución conocido

- Como esta variable no se observa, sino que se observa y, entonces se puede deducir que se observa $y = 1$ si $y^* > 0$ y que $y = 0$ si $y^* \leq 0$. Esto se suele denotar por $\mathbf{1}(y^* > 0)$. En esta formulación, $\boldsymbol{x}'\boldsymbol{\beta}$ se denomina la función índice

- La suposición de varianza conocida de ε es una normalización, pero si se considera un parámetro σ_ε^2 no restringido, entonces la regresión latente será $y^* = \boldsymbol{x}'\boldsymbol{\beta} + \sigma\varepsilon^*$, donde ahora ε tiene una varianza de 1

  - No obstante, $(y^*/\sigma) = \boldsymbol{x}'(\boldsymbol{\beta}/\sigma) + \varepsilon$ es el mismo modelo con los mismos datos. Los datos observados y no cambian, dependiendo solo del signo de y^* y no de su escala como se puede ver

  - Esto significa que no hay información sobre σ en los datos muestrales y que, por tanto, no se puede estimar y el vector de parámetros $\boldsymbol{\beta}$ solo se identifica hasta la escala

- La suposición de que el umbral para la observación es nulo tampoco es realista si el término constante se incluye en la regresión (aunque si lo es si este no se incluye)

  - Siendo a un umbral diferente de cero y α un término constante desconocido para la regresión, y pensando que $\boldsymbol{x}$ y $\boldsymbol{\beta}$ incluyen el resto del índice (sin tener un término constante), entonces la probabilidad de y se puede expresar de la siguiente manera:

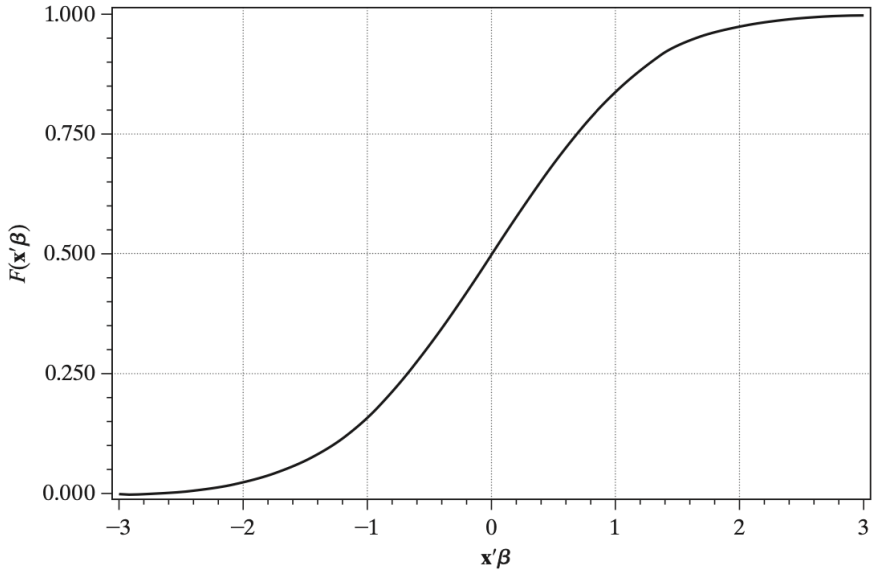
$$P(y^* > a|\boldsymbol{x}) = P(\alpha + \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon > a|\boldsymbol{x}) =$$

$$= P\big((\alpha - a) + \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon > 0\big|\boldsymbol{x}\big)$$

  - Debido a que α es desconocido, entonces la diferencia $(\alpha - a)$ es un parámetro desconocido. El resultado final es que, si la regresión contiene un término constante, entonces este no cambia por la elección del umbral, por lo que fijarlo en cero es una normalización sin importancia. Esto permite obtener la siguiente igualdad:

$$P(y^* > 0|\boldsymbol{x}) = P(\varepsilon > -\boldsymbol{x}'\boldsymbol{\beta}|\boldsymbol{x})$$

- Dependiendo del tipo de distribución que se asume para el modelo, entonces se puede obtener una forma funcional para la función de distribución de probabilidad para ε



  - Se supone que un conjunto de factores del vector $\boldsymbol{x}$ explica la decisión del individuo, mientras que $\boldsymbol{\beta}$ es un vector de parámetros que refleja el impacto de cambios en $\boldsymbol{x}$ en la probabilidad. Por lo tanto, solo se requiere una especificación adecuada para la parte derecha de la ecuación

$$P(Y = 1|\boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta}) \qquad P(Y = 0|\boldsymbol{x}) = 1 - F(\boldsymbol{x}, \boldsymbol{\beta})$$

  - El único requerimiento es que se produzcan predicciones consistentes con la teoría subyacente vista anteriormente. Por lo tanto, para un vector de regresores, se espera que se cumplan las siguientes condiciones:

$$0 \leq P(Y = 1|\boldsymbol{x}) \leq 1$$

$$\lim_{\boldsymbol{x'\beta} \to -\infty} P(Y = 1|\boldsymbol{x}) = 0 \qquad \lim_{\boldsymbol{x'\beta} \to \infty} P(Y = 1|\boldsymbol{x}) = 1$$

  - En principio, cualquier distribución definida sobre la recta real es válida, pero si esta distribución es simétrica (como la logística o la normal), entonces se puede obtener la siguiente igualdad:

$$P(y^* > 0|\boldsymbol{x}) = P(\varepsilon > -\boldsymbol{x'\beta}|\boldsymbol{x}) = P(\varepsilon < \boldsymbol{x'\beta}|\boldsymbol{x}) = F(\boldsymbol{x'\beta})$$

- Algunas de las distribuciones más utilizadas son la normal, la logística o la distribución de Gumbel, las cuales dan paso a diferentes modelos. No obstante, los marcos más utilizados son los dos primeros

- Si se utiliza la distribución normal, se da paso al modelo probit (con distribución simétrica):

$$P(Y = 1|\boldsymbol{x}) = \int_{-\infty}^{\boldsymbol{x}'\boldsymbol{\beta}} \phi(t)\, dt = \Phi(\boldsymbol{x}'\boldsymbol{\beta})$$

- Si se utiliza la distribución logística, se da paso al modelo logit (con distribución simétrica):

$$P(Y = 1|\boldsymbol{x}) = \int_{-\infty}^{\boldsymbol{x}'\boldsymbol{\beta}} \Lambda(t)[1 - \Lambda(t)]\, dt = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}'\boldsymbol{\beta})} = \Lambda(\boldsymbol{x}'\boldsymbol{\beta})$$

- Si se utiliza la distribución de Gumbel o el modelo de valores extremos de tipo I, se obtiene el siguiente modelo:

$$P(Y = 1|\boldsymbol{x}) = \exp\left[-\exp\left(-\boldsymbol{x}'\boldsymbol{\beta}\right)\right]$$

- El modelo de log-log complementario resta el modelo de Gumbel a la unidad:

$$P(Y = 1|\boldsymbol{x}) = 1 - \exp\left[-\exp\left(-\boldsymbol{x}'\boldsymbol{\beta}\right)\right]$$

- El modelo de Burr es parecido al logit, solo que la razón resultante se eleva a un parámetro γ

$$P(Y = 1|\boldsymbol{x}) = \left[\frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}'\boldsymbol{\beta})}\right]^{\gamma} = [\Lambda(\boldsymbol{x}'\boldsymbol{\beta})]^{\gamma}$$

- Las colas de la distribución logística son más gruesas que la de la distribución normal, de modo que para valores intermedios se pueden obtener probabilidades similares, mientras que para valores en las colas se obtienen probabilidades mucho más grandes para la distribución logística que para la distribución normal

- La mayoría de análisis se basan en examinar las relaciones entre los regresores $\boldsymbol{x}$ y la probabilidad del evento $P(Y = 1|\boldsymbol{x}) = F(y|\boldsymbol{x}) = F(\boldsymbol{x}'\boldsymbol{\beta})$, las cuales se interpretan a través de los efectos parciales. Como el modelo no es lineal, los parámetros $\boldsymbol{\beta}$ no son necesariamente los efectos parciales de los modelos

$$\frac{\partial F(y|\boldsymbol{x})}{\partial \boldsymbol{x}} = \left[\frac{\partial F(\boldsymbol{x}'\boldsymbol{\beta})}{\partial(\boldsymbol{x}'\boldsymbol{\beta})}\right]\boldsymbol{\beta} = f(\boldsymbol{x}'\boldsymbol{\beta})\boldsymbol{\beta}$$

- En el caso de la distribución normal y de la distribución logística se obtienen los siguientes resultados:

$$\frac{\partial F(y|\boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{\partial \Phi(\boldsymbol{x}'\boldsymbol{\beta})}{\partial(\boldsymbol{x}'\boldsymbol{\beta})} = \phi\left(\boldsymbol{x}'\boldsymbol{\beta}_{\boldsymbol{prob}}\right)\boldsymbol{\beta}_{\boldsymbol{prob}}$$

$$\frac{\partial F(y|\boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{\partial \Lambda(\boldsymbol{x}'\boldsymbol{\beta})}{\partial(\boldsymbol{x}'\boldsymbol{\beta})} = \Lambda\left(\boldsymbol{x}'\boldsymbol{\beta}_{\boldsymbol{log}}\right)\left[1 - \Lambda\left(\boldsymbol{x}'\boldsymbol{\beta}_{\boldsymbol{log}}\right)\right]\boldsymbol{\beta}_{\boldsymbol{log}}$$

- Como se puede ver, los efectos dependerán de los valores de $\boldsymbol{x}$. En los modelos de función de índice, generalmente, el vector de efectos parciales es un múltiplo del vector de coeficientes

- En muchas aplicaciones, una regularidad empírica que suele aparecer es que $1.6\widehat{\boldsymbol{\beta}}_{\boldsymbol{prob}} \approx \widehat{\boldsymbol{\beta}}_{\boldsymbol{log}}$, lo cual puede sugerir una gran diferencia entre ambos modelos. Esto ocurre porque el efecto parcial del modelo *probit* para el valor medio sería $0.4\widehat{\boldsymbol{\beta}}_{\boldsymbol{prob}}$, mientras que para el modelo *logit* el valor sería $0.5(1 - 0.5)\widehat{\boldsymbol{\beta}}_{\boldsymbol{prob}} = 0.25\widehat{\boldsymbol{\beta}}_{\boldsymbol{log}}$, de modo que, para el mismo valor medio, se obtiene que $\widehat{\boldsymbol{\beta}}_{\boldsymbol{log}} = (0.4/0.25)\widehat{\boldsymbol{\beta}}_{\boldsymbol{prob}}$

- Existen varias maneras de poder obtener inferencias a partir de estimadores del efecto parcial

  - Para calcular los efectos parciales, normalmente se puede evaluar las expresiones en las medias muestrales de los datos, produciendo efectos parciales en las medias o *partial effects at averages* (PEA)

$$PEA = \widehat{\boldsymbol{\gamma}}(\overline{\boldsymbol{x}}) = f\left(\overline{\boldsymbol{x}}'\widehat{\boldsymbol{\beta}}\right)\widehat{\boldsymbol{\beta}}$$

  - Como las medias no siempre producen un escenario realista, es mejor calcular los efectos parciales en cada punto y obtener la media muestral de los efectos parciales individuales, produciendo los efectos parciales medios o *average partial effects*

$$APE = \widehat{\overline{\boldsymbol{\gamma}}} = \frac{1}{n}\sum_{i=1}^{n} f\left(\boldsymbol{x}'\widehat{\boldsymbol{\beta}}\right)\widehat{\boldsymbol{\beta}}$$

  - Normalmente el interés reside en el efecto parcial medio o *average partial effect*, que es el valor esperado del efecto parcial

$$APE^0 = \boldsymbol{\gamma}^0 = E_x\left[\frac{\partial E(y|\boldsymbol{x})}{\partial \boldsymbol{x}}\right]$$

- Se pueden interpretar los parámetros del modelo utilizando diferentes enfoques

  - Debido a que los parámetros no se pueden interpretar directamente como efectos parciales, solo se puede interpretar el signo de la relación con la probabilidad, ya que el signo del coeficiente determinará el signo del efecto parcial

$$\beta_k < 0 \;\Rightarrow \frac{\partial F(y|\boldsymbol{x})}{\partial x_k} < 0 \qquad \beta_k > 0 \;\Rightarrow \frac{\partial F(y|\boldsymbol{x})}{\partial x_k} > 0$$

  - Para una variable binaria cualquiera, es común calcular la diferencia de probabilidad con el caso base para analizar el efecto de pasar de una categoría a otra en términos de probabilidad

$$P(y|\boldsymbol{x}, d = 1) - P(y|\boldsymbol{x}, d = 0) = F(y|\boldsymbol{x}, d = 1) - F(y|\boldsymbol{x}, d = 0)$$

  - También se puede comparar la importancia relativa del efecto parcial de una variable con el efecto parcial del de otra

$$\frac{\frac{\partial F(y|\boldsymbol{x})}{\partial x_k}}{\frac{\partial F(y|\boldsymbol{x})}{\partial x_l}} = \frac{f(\boldsymbol{x}'\boldsymbol{\beta})\beta_k}{f(\boldsymbol{x}'\boldsymbol{\beta})\beta_l} = \frac{\beta_k}{\beta_l}$$

- Una medida que ayuda a interpretar el resultado del modelo es la *odds ratio*, el cual se basa en la medida de *odds in favour*, la cual es una razón $P(Y = 1|\boldsymbol{x})/P(Y = 0|\boldsymbol{x})$ que expresa las probabilidades a favor de que ocurra un evento en el modelo *logit*

$$Odds = \frac{P(Y = 1|\boldsymbol{x})}{P(Y = 0|\boldsymbol{x})} = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})\,/[1 + \exp(\boldsymbol{x}'\boldsymbol{\beta})]}{1/[1 + \exp(\boldsymbol{x}'\boldsymbol{\beta})]} = \exp(\boldsymbol{x}'\boldsymbol{\beta})$$

  - Considerando el efecto en las *odds* cuando cambia una variable binaria d, se obtiene la *odds ratio* (la razón entre las *odds* cuando se incrementa un regresor en una unidad y las *odds* del caso base):

$$\frac{Odds(\boldsymbol{x}, d = 1)}{Odds(\boldsymbol{x}, d = 0)} = \frac{\frac{\exp(\boldsymbol{x}'\boldsymbol{\beta} + \delta)\,/[1 + \exp(\boldsymbol{x}'\boldsymbol{\beta} + \delta)]}{1/[1 + \exp(\boldsymbol{x}'\boldsymbol{\beta} + \delta)]}}{\frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})\,/[1 + \exp(\boldsymbol{x}'\boldsymbol{\beta})]}{1/[1 + \exp(\boldsymbol{x}'\boldsymbol{\beta})]}} = \exp(\delta)$$

  - Por lo tanto, cuando una variable cambia una unidad, la *odds ratio* permite aproximar el efecto parcial de la variable, aunque no es la derivada. Es decir, el exponencial del coeficiente β_k representa el cambio multiplicativo en la *ratio* y $\exp(\beta_k) - 1$,

por tanto, representa el cambio proporcional al aumentar en una unidad la variable x_k

- El resultado binario sugiere un modelo de regresión $F(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{x}'\boldsymbol{\beta}$ con esperanza condicional $E(y|\boldsymbol{x}) = 0[1 - F(\boldsymbol{x}, \boldsymbol{\beta})] + F(\boldsymbol{x}, \boldsymbol{\beta}) = F(\boldsymbol{x}, \boldsymbol{\beta})$, lo cual implica el siguiente modelo de regresión:

$$y = E(y|\boldsymbol{x}) + [y - E(y|\boldsymbol{x})] = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon$$

  - A este modelo se le llama el modelo de probabilidad lineal, aunque este tiene algunos problemas

  - En este modelo, ε es heteroscedástico y depende de $\boldsymbol{\beta}$, lo cual se debe a la naturaleza binaria de la variable dependiente. Esto se puede ver fácilmente con la varianza, y se puede gestionar utilizando el estimador de mínimos cuadrados generalizado factible (aunque no soluciona el problema teórico, solo el de estimación)

$$\begin{cases} y_i = 1 \quad if \quad \varepsilon_i = 1 - \boldsymbol{x}'\boldsymbol{\beta} \\ y_i = 0 \quad if \quad \varepsilon_i = -\boldsymbol{x}'\boldsymbol{\beta} \end{cases}$$

$$Var(\varepsilon|\boldsymbol{x}) = (1 - \boldsymbol{x}'\boldsymbol{\beta})^2 F(\boldsymbol{x}, \boldsymbol{\beta}) + (-\boldsymbol{x}'\boldsymbol{\beta})^2 \big(1 - F(\boldsymbol{x}, \boldsymbol{\beta})\big) =$$

$$= (1 - \boldsymbol{x}'\boldsymbol{\beta})^2 \boldsymbol{x}'\boldsymbol{\beta} + (-\boldsymbol{x}'\boldsymbol{\beta})^2 (1 - \boldsymbol{x}'\boldsymbol{\beta}) = \boldsymbol{x}'\boldsymbol{\beta}(1 - \boldsymbol{x}'\boldsymbol{\beta})$$

  - Un problema más grave es que no se garantiza que el valor de y esté acotado entre cero y uno (que se comporten como probabilidades). No obstante, a veces se utiliza porque permite obtener efectos parciales parecidos a los obtenidos con modelos *probit* y *logit* y por la robustez del modelo a suposiciones distribucionales

- Exceptuando el modelo de probabilidad lineal, la estimación de los modelos de elección binaria se lleva a cabo a través del método de máxima verosimilitud

  - Cada observación se trata como una realización de una distribución de Bernoulli, en donde la probabilidad de éxito es $F(\boldsymbol{x}'\boldsymbol{\beta})$ y las observaciones son independientes. Por lo tanto, la función de verosimilitud es la siguiente:

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n|\boldsymbol{X}) = \prod_{y_i=0} [1 - F(\boldsymbol{x_i}'\boldsymbol{\beta})] \prod_{y_i=1} F(\boldsymbol{x_i}'\boldsymbol{\beta})$$

    - Esta se puede reexpresar de manera conveniente para una muestra de n observaciones y para el logaritmo de la función

$$L(\boldsymbol{\beta}|data) = \prod_{i=0}^{n}[1 - F(\boldsymbol{x_i'\beta})]^{y_i}[F(\boldsymbol{x_i'\beta})]^{1-y_i}$$

$$\ln L = \sum_{i=0}^{n}\left[y_i \ln F(\boldsymbol{x_i'\beta}) + (1 - y_i)\ln\left(1 - F(\boldsymbol{x_i'\beta})\right)\right]$$

- Por lo tanto, el estimador para $\boldsymbol{\beta}$ será aquel que satisfazca la siguiente ecuación:

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=0}^{n}\left[y_i \frac{f(\boldsymbol{x_i'\beta})}{F(\boldsymbol{x_i'\beta})} + (1 - y_i)\frac{-f(\boldsymbol{x_i'\beta})}{1 - F(\boldsymbol{x_i'\beta})}\right]\boldsymbol{x_i} = \mathbf{0}$$

  - Mientras que no se esté en el modelo lineal, la ecuación será no lineal y serán necesarios métodos iterativos

  - Para el modelo *logit* se obtiene la siguiente condición de primer orden:

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=0}^{n}[y_i - \Lambda_i]\boldsymbol{x_i} = \mathbf{0}$$

  - Si $\boldsymbol{x_i}$ contiene un término constante, entonces la media muestral de las probabilidades predichas debe ser igual a la proporción de unos en la muestra. Además, esta implicación tiene similitudes con las ecuaciones normales de mínimos cuadrados si se interpreta $y_i - \Lambda_i$ como el error

  - Para el modelo *probit* se obtiene la siguiente condición de primer orden:

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{y_i=0}\frac{-\phi_i}{1 - \Phi_i}\boldsymbol{x_i} + \sum_{y_i=1}\frac{\phi_i}{\Phi_i}\boldsymbol{x_i} = \sum_{y_i=0}\lambda_{0i}\boldsymbol{x_i} + \sum_{y_i=1}\lambda_{1i}\boldsymbol{x_i} = \mathbf{0}$$

- Se han sugerido varias medidas para modelos de respuesta discreta que intentan imitar al R^2, ya que mide el éxito de un estimador al optimizar el criterio de ajuste y, si se calcula como ρ^2, mide cómo las predicciones del modelo pueden imitar los datos reales

  - Debido a que muchos modelos se estiman por máxima verosimilitud, de modo que el criterio de ajuste es la verosimilitud

    - Siguiendo con el criterio del R^2, es interesante ver la hipótesis de que todas las pendientes sean nulas. La log-verosimilitud en el caso de que solo haya un término constante sería:

$$\ln L_0 = n[P_0 \ln P_0 + P_1 \ln P_1] \quad where \;\; P_j \;\; is\, proportion$$

- El indicador más usado es la pseudo-R^2 es una medida que está acotada entre cero y uno, pero que no puede llegar nunca a uno (la verosimilitud nunca puede ser nula)

$$R^2_{pseudo} = 1 - \frac{\ln L_{MLE}}{\ln L_0}$$

- Es posible hacer una corrección de grados de libertad utilizando el siguiente estadístico:

$$R^2_{pseudo} = 1 - \frac{\ln L_{MLE} - K}{\ln L_0}$$

- El pseudo-R^2 es muy útil para comparar un modelo con otro, pero si los modelos son anidados, entonces la función de log-verosimilitud es la elección natural. Para casos más generales, los investigadores utilizan criterios de información

  - Los dos criterios de información más utilizados son el criterio de información de Akaike y el criterio de información de Bayes

$$AIC = -2 \ln L + 2K \quad or \quad AIC/n$$

$$BIC = -2 \ln L + K \ln n \quad or \quad BIC/n$$

  - En general, cuanto menor sea el valor del criterio mejor será el ajuste

Los modelos de elección discreta multinomial desordenada

- Cuando la variable dependiente toma valores discretos, pero no es binaria (no toma solo dos valores, sino que toma tres o más), el modelo lineal sigue siendo un modelo inadecuado para estudiar la elección discreta

  - Los modelos anteriores se pueden generalizar con tal estudiar las probabilidades de que un individuo seleccione una alternativa j respecto a todas las alternativas posibles J, de modo que los números de la variable representan diferentes alternativas

    - Se suele utilizar el modelo *logit* multinomial y el modelo *logit* condicional

- Los coeficientes de los regresores $\boldsymbol{x}$ siguen siendo no interpretables, de modo que se sigue utilizando el concepto de efecto marginal

- Se sigue usando el modelo de utilidad aleatoria o *random utility model*, pero generalizado al caso en el que hay más de dos alternativas. En este caso, los coeficientes que expresan la diferencia entre un grupo base y otro grupo (el vector β) variarán dependiendo de la alternativa (a diferencia del caso anterior):

$$U_{ij} = \boldsymbol{x}_i'\boldsymbol{\gamma_j} + u_{ij} \quad for \;\; j = 1,2,\dots,J \;\; \Rightarrow \;\; U_{ij} - U_{ig} = \boldsymbol{x}_i'\boldsymbol{\beta_j} + \varepsilon_{ij}$$

$$where \;\; \boldsymbol{\beta_j} = \boldsymbol{\gamma_j} - \boldsymbol{\gamma_g} \;\; for \;\; j = 0,1,2,\dots,J$$

  - Cuando la variable dependiente es equivalente a una alternativa j, quiere decir que, según el modelo, la utilidad de la alternativa j es mayor a la de la alternativa g (se sigue el modelo de maximización de utilidad del consumidor)

$$U_{ij} = \max(U_{i1}, U_{i2}, \dots)$$

$$Y_i = j \quad if \quad U_{ij} > U_{ig} \;\; when \;\; j \neq g \;\; for \;\; j = 0,1,2,\dots,J$$

  - En este caso la probabilidad de que se escoja la alternativa j (que $Y_i = j$) se generaliza al caso en el que el grupo base y el grupo de la alternativa no son iguales:

$$P(Y_i = j|\boldsymbol{x_i}) = P\big(U_{ij} > U_{ig}\big|\boldsymbol{x_i}\big) = P\big(U_{ij} - U_{ig} > 0\big|\boldsymbol{x_i}\big)$$

$$= P\big(\varepsilon_{ig} > \boldsymbol{x_i'\beta_j} + \varepsilon_{ij}\big|\boldsymbol{x_i}\big) \;\; for\, all\, j \neq g$$

  - Se asume que el término de error para cada alternativa es idéntico e independiente del término de las otras alternativas, y que para los casos multinomiales sigue una distribución de Gumbell o *extreme value type I*

$$\boldsymbol{u_i}\; is\; i.i.d. \;\; \Rightarrow \;\; \boldsymbol{\varepsilon_i}\; is\; i.i.d.$$

$$F\big(\varepsilon_{ij}\big) = \exp\left(-\frac{1}{e^{\varepsilon_{ij}}}\right) \;\Rightarrow\; u_{ij}, \varepsilon_{ij} \sim Gumbel$$

$$where \;\; \boldsymbol{u_i} = \begin{pmatrix} u_{i1} \\ u_{i2} \\ \dots \end{pmatrix} \; and \; \boldsymbol{\varepsilon_i} = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \dots \end{pmatrix}$$

  - Si se cree que el error se distribuye de manera normal (con media cero y varianza constante), en vez de seguir una

distribución *Extreme Value Type-I,* se utiliza el modelo *probit* multinomial

- El modelo *logit* multinomial es un modelo generalizado del modelo *logit*. Es un modelo de regresión no lineal que utiliza la función de distribución de probabilidad acumulada de la distribución logística, con la que, a partir del *additive random utility model* se demuestra que la probabilidad de que $Y = j$ (de entre J alternativas mutuamente excluyentes) es:

$$E(Y_i = j|\boldsymbol{x_i}) = P(Y_i = j|\boldsymbol{x_i}) = \frac{e^{(\boldsymbol{x_i'\beta_j})}}{\sum_{m=1}^{J} e^{(\boldsymbol{x_i'\beta_m})}}$$

  - En el modelo, igual que en el caso binario, se escoge una alternativa como grupo o categoría base y, por tanto, los coeficientes del modelo pueden ser interpretados respecto a la base como la diferencia entre escoger la alternativa j o la base

    - Esto implica que los coeficientes de la categoría base serán todos nulos porque si no habría un problema de identificación (demasiados coeficientes)

$$\boldsymbol{\beta_g} = 0$$

    - Si los coeficientes de la alternativa base son nulos (por suposición), la probabilidad de escoger la alternativa base serían:

$$P(Y_i = g|\boldsymbol{x_i}) = \frac{1}{\sum_{m=1}^{J} e^{(\boldsymbol{x_i'\beta_m})}}$$

  - Hay dos propiedades que se cumplen en los modelos *logit* si se incluye el coeficiente constante:

    - Como las alternativas tienen que ser mutuamente excluyentes, la suma de las probabilidades de escoger cada alternativa para cada individuo debe ser igual a la unidad (con certeza escogerá una alternativa). Esto implica que el cambio en la probabilidad de una alternativa hace que cambie la probabilidad de las otras

$$\sum_{j=1}^{J} P(Y_i = j|\boldsymbol{x_i}) = 1 \quad for \quad i = 1,2,\dots,N$$

$$P(Y_i = j|\boldsymbol{x_i}) \uparrow \;\Rightarrow\; P(Y_i = m|\boldsymbol{x_i}) \downarrow \quad when\, j \neq m$$

    - Igual que en el caso binario, si se incluye el coeficiente constante, la suma de probabilidades ajustadas del modelo *logit* coincide con el número de individuos con valor $Y = j$

$$\sum_{i=1}^{N} P(Y_i = j|\boldsymbol{x_i}) = N_j \quad for \quad j = 1,2,\dots,J$$

- Si se no se incluye el coeficiente constante, estas propiedades dejan de ser ciertas

- Este modelo se puede modelar a través de variables binarias (respuesta) para cada una de las J categorías

$$D_{ij} = \begin{cases} 1 & if \ \ Y_i = j \\ 0 & if \ \ Y_i \neq j \end{cases} \quad for \ \ j = 0,1,\dots,J-1$$

- Se crean $J-1$ variables binarias porque la categoría base no se modelaría

- De este modo, en vez de tener un solo modelo, se tienen J modelos de regresión binaria *logit*, dado que se modela por separado y no se cuenta el orden

$$p_{ij} = P\left(D_{ij} = 1 \middle| \boldsymbol{x_i}\right) = P(Y_i = j|\boldsymbol{x_i})$$

- Igual que en los modelos *logit* y *probit* binomiales, en el modelo *logit* multinomial no se pueden interpretar directamente los coeficientes, de modo que se utiliza el efecto marginal

  - En este caso, los efectos marginales dependen no solo del individuo y del regresor, si no también de la alternativa, debido a que cada alternativa tiene un coeficiente. Además, también depende de las probabilidades de las otras alternativas:

$$ME_{ikj} = \frac{\partial P(Y_i = j|\boldsymbol{x_i})}{\partial X_{ik}}$$

$$\frac{\partial P(Y_i = j|\boldsymbol{x_i})}{\partial X_{ikj}} = P(Y_i = j|\boldsymbol{x_i}) \left[\beta_{jk} - \sum_{m=1}^{J} \beta_{mk} P(Y_i = m|\boldsymbol{x_i}) \right]$$

- La sumatoria de las probabilidades de cada alternativa multiplicadas por el coeficiente de un regresor k (para cada alternativa) se puede interpretar como una media ponderada de los coeficientes del regresor k a través de las alternativas. De este modo, cada coeficiente de X_{ik} entra en el efecto marginal de cada alternativa

$$\bar{\beta}_k = \sum_{m=1}^{J} \beta_{mk} P(Y_i = m|\boldsymbol{x_i})$$

- Además, la diferencia entre el coeficiente y la media ponderada no tiene por qué tener el mismo signo que el coeficiente β_{jk}, por lo que no se puede interpretar los signos de los coeficientes (como sí que ocurría en el modelo *logit* binomial)

- Como β_{jk} es uno de los coeficientes que se tienen en cuenta también al hacer la media ponderada $\bar{\beta}_k$, solo se sabe con certeza que el coeficiente con el valor máximo es mayor a la media y que aquel con valor mínimo es menor:

$$\beta_{ak} < \sum_{m=1}^{J} P(Y_i = m|\boldsymbol{x_i})\beta_{mk} < \beta_{bk}$$

$$a = alternative\ with\ minimum\ \beta\ value$$

$$b = alternative\ with\ maximum\ \beta\ value$$

- Para la alternativa con el coeficiente con un valor máximo o mínimo, por tanto, se puede saber el signo del efecto marginal, dado que se sabe el signo de la diferencia

$$\beta_{bk} - \sum_{m=1}^{J} P(Y_i = m|\boldsymbol{x_i})\beta_{mk} > 0 \ \Rightarrow \ ME_{ikb} > 0$$

$$\beta_{ak} - \sum_{m=1}^{J} P(Y_i = m|\boldsymbol{x_i})\beta_{mk} < 0 \ \Rightarrow \ ME_{ika} < 0$$

- Los coeficientes en el modelo *logit* multinomial también se estiman mediante la función de máxima verosimilitud, la cual se generaliza en este modelo:

$$L = \prod_{i=1}^{N_1} P(Y_i = 1|\boldsymbol{x_i}) \ \ldots \prod_{i=N_{J-1}+1}^{N_J} P(Y_i = J|\boldsymbol{x_i}) = \prod_{j=1}^{J} \prod_{i=N_{j-1}+1}^{N_j} P(Y_i = j|\boldsymbol{x_i})$$

$$\max_{\boldsymbol{\beta_j}} \ln\left[\prod_{j=1}^{J} \prod_{i=N_{j-1}+1}^{N_j} P(Y_i = j|\boldsymbol{x_i})\right] = \sum_{j=1}^{J} \sum_{j=N_{j-1}+1}^{N_j} \ln\left[\frac{e^{(\boldsymbol{x_i'\beta_j})}}{\sum_{r=1}^{J} e_j^{(\boldsymbol{x_i'\beta_j})}}\right]$$

- Existen dos casos especiales que simplifican la obtención de los coeficientes maximizadores:

- Cuando solo se tiene que estimar el coeficiente constante es como si se tuviera que estimar la media, y el coeficiente constante que maximiza la función de verosimilitud será aquel que haga que la probabilidad sea la proporción de observaciones en la muestra donde $Y = j$:

$$p_j^* = P^*\left(Y_i = j|\boldsymbol{x_{ij}}\right) = \frac{e^{\beta_{0j}}}{\sum_{m=1}^{J} e^{\beta_{0m}}} = \frac{N_j}{N} = \frac{\sum y_i}{N} = \overline{Y_j}$$

$$L = \left(\frac{N_1}{N}\right)^{N_1} \dots \left(\frac{N_J}{N}\right)^{N_J} = \prod_{j=1}^{J} \left(\frac{N_j}{N}\right)^{N_j}$$

- Cuando se incluye un solo regresor binario, los individuos se separan en dos grupos, lo cual hace que los coeficientes maximizadores sean aquellos que hagan que la probabilidad de que $Y = 1$ sea igual a la proporción de individuos en la muestra con ese valor para cada grupo:

$$P(Y_i = j|X_i = 1) = \frac{e^{(\beta_{0j}+\beta_{1j})}}{\sum_{m=1}^{J} e^{(\beta_{0m}+\beta_{1m})}} \Rightarrow P^*(Y_i = j|X_i = 1) = \frac{N_{j,X=1}}{N_{X=1}}$$

$$P(Y = j|X_i = 0) = \frac{e^{\beta_{0j}}}{\sum_{m=1}^{J} e^{\beta_{0m}}} \Rightarrow P^*(Y = j|X_i = 0) = \frac{N_{j,X=0}}{N_{X=0}}$$

$$L = \left(\frac{N_{1,X=1}}{N_{X=1}}\right)^{N_{1,X=1}} \left(\frac{N_{2,X=1}}{N_{X=1}}\right)^{N_{2,X=1}} \dots \left(\frac{N_{1,X=0}}{N_{X=0}}\right)^{N_{1,X=0}} \left(\frac{N_{2,X=0}}{N_{X=0}}\right)^{N_{2,X=0}} \dots$$

$$L = \prod_{j=1}^{J} \left(\frac{N_{j,X=1}}{N_{X=1}}\right)^{N_{j,X=1}} \prod_{j=1}^{J} \left(\frac{N_{j,X=0}}{N_{X=0}}\right)^{N_{j,X=0}}$$

- Este último caso particular puede generalizarse para varias variables binarias

- Cuando se produce un cambio de categoría o grupo base y se tiene el valor de los coeficientes estimados, se pueden estimar los nuevos coeficientes porque se mantienen las diferencias entre coeficientes:

$$\beta_{1j} - \beta_{2j} = \beta'_{1j} - \beta'_{2j}$$

- Esto ocurre porque, cuando se cambia de base, las probabilidades de un individuo i para las J alternativas no tiene que variar. En consecuencia, tampoco varía la *odds ratio* (la probabilidad relativa de una alternativa respecto a otra para un individuo i)

- Tampoco cambian los efectos marginales, debido a que no cambian las probabilidades

- Otro modelo que se utiliza es el *logit* condicional, el cual es una ampliación del modelo *logit* multinomial. En este modelo, se continúan teniendo J alternativas, pero ahora hay dos tipos de regresores: unos que no dependen de la alternativa (las $\boldsymbol{X}$, pero varían según el individuo i) y otros que sí lo hacen (por lo que estos varían según la alternativa j para un mismo individuo i)

$$U_{ij} = \boldsymbol{x}_i'\boldsymbol{\gamma_j} + \boldsymbol{z}_{ij}'\boldsymbol{\alpha} + u_{ij}$$

$$k = number\ of\ regressor\ independent\ of\ j$$

  - En este modelo, los coeficientes de los regresores que varían según la alternativa son constantes, de modo que en la diferencia de utilidades solo se tiene en cuenta la diferencia del regresor entre alternativas:

$$U_{ij} - U_{ig} = \boldsymbol{x}_i^{'}\left(\boldsymbol{\gamma_j} - \boldsymbol{\gamma_g}\right) + (\boldsymbol{z}_{ij}^{'} - \boldsymbol{z}_{ig}^{'})\boldsymbol{\alpha} + (u_{ij} - u_{ig})$$

$$U_{ij} - U_{ig} = \boldsymbol{x}_i^{'}\boldsymbol{\beta_j} + (\boldsymbol{z}_{ij}^{'} - \boldsymbol{z}_{ig}^{'})\boldsymbol{\alpha} + \varepsilon_{ij}$$

    - Como los coeficientes $\boldsymbol{\alpha}$ siempre son los mismos, estos no necesitan una alternativa base para ser calculados (si no se tuvieran coeficientes constantes habría un problema de mala especificación). Sin embargo, aún se necesita seleccionar una alternativa base para la estimación de los coeficientes $\boldsymbol{\beta}$

    - Los errores siguen teniendo la misma distribución asumida anteriormente

  - De este modo, la probabilidad de escoger la probabilidad j será:

$$P\left(Y_i = j \middle| \boldsymbol{x_{ij}}, \boldsymbol{z_i}\right) = \frac{e^{(\boldsymbol{x_i'\beta_j} + \boldsymbol{\alpha z_{ij}'})}}{\sum_{m=1}^{J} e^{(\boldsymbol{x_i'\beta_m} + \boldsymbol{\alpha z_{im}'})}}$$

$$g = reference\ category$$

    - En este caso, la diferencia entre $\boldsymbol{z}$ para cada alternativa no se cuenta, dado que cuando se estima el modelo, $\boldsymbol{\alpha}$ representa el cambio en $U_{ij} - U_{ig}$ cuando $\boldsymbol{z_{ij}}$ varía una unidad

  - En el modelo *logit* condicional, los coeficientes de los regresores invariantes con la alternativa (las $\boldsymbol{\beta}$) se interpretan del mismo modo que en el modelo *logit* multinomial, pero los coeficientes de los regresores que varían según la alternativa (las $\boldsymbol{\alpha}$) se interpretan como en el modelo *logit* binomial

- Como los coeficientes $\boldsymbol{\beta}$ dependen de la alternativa j, solo se puede interpretar aquellos que tengan el valor máximo y el mínimo (a través de sus efectos marginales)

$$\frac{\partial P(Y_i = j|\boldsymbol{x_i}, \boldsymbol{z_{ij}})}{\partial X_{ik}} = P(Y_i = j|\boldsymbol{x_i}, \boldsymbol{z_{ij}})(\beta_{jk} - \bar{\beta}_k)$$

- Como los coeficientes $\boldsymbol{\alpha}$ no varían con la alternativa (siempre son los mismos para cualquier alternativa j), el efecto marginal de variar una unidad de un regresor Z_{ij} tendrá el mismo signo que su coeficiente α, pero la magnitud sigue siendo no interpretable

$$\frac{\partial P(Y_i = j|\boldsymbol{x_i}, \boldsymbol{z_{ij}})}{\partial Z_{ikj}} = \left[P(Y_i = j|\boldsymbol{x_i}, \boldsymbol{z_{ij}})[\mathbf{1}(j = m) - P(Y_i = m|\boldsymbol{x_i}, \boldsymbol{z_{ij}})]\right]\alpha_k$$

$$where \quad \mathbf{1}(j = m) = \begin{cases} 1 \ \ if\ j = m \\ 0 \ \ if\ j \neq m \end{cases}$$

  - Cuando incrementa una unidad del regresor Z_{ij}, el efecto en la probabilidad del individuo i de escoger la alternativa j tendrá el mismo signo. Sin embargo, el efecto en la probabilidad de que el individuo i escoja las otras alternativas tendrá un signo contrario, debido a que el individuo i tiene que escoger una alternativa de las posibles

$$\sum_{m=1}^{J} P(Y_i = j|\boldsymbol{x}_i, \boldsymbol{z}_{im}) = 1 \quad \Rightarrow \begin{cases} P(Y_i = 1|\boldsymbol{x}_i, \boldsymbol{z}_{ij}) \uparrow \\ P(Y_i = 2|\boldsymbol{x}_i, \boldsymbol{z}_{ij}) \downarrow \\ \dots \downarrow \end{cases}$$

    - A diferencia de con variaciones en $\boldsymbol{X}_i$, en donde el efecto del cambio depende de la alternativa j en la que se esté, el efecto de $\boldsymbol{Z_{ij}}$ es el mismo para cualquier alternativa, de modo que el efecto de la variación en los regresores $\boldsymbol{Z_{ij}}$ para j es opuesto para las otras alternativas $m \neq j$

- Los modelos *logit* multinomiales y *logit* condicionales cumplen una propiedad conveniente para la estimación, pero que no siempre se adecua a la realidad: la independencia de las alternativas irrelavantes o *independence of irrelevant alternatives* (IIA)

  - La *odds ratio* es la probabilidad relativa entre una alternativa j y una alternativa base g, la cual se define como la *ratio* entre ambas

    - Para el caso de un modelo *logit* multinomial, se puede ver como la *odds ratio* solo depende de j y r, por lo que las otras alternativas son irrelevantes y los coeficientes estimados para

las otras alternativas no tendrían que variar nada o por lo menos no mucho si se extraen alternativas

$$\frac{P(Y_i = j|\boldsymbol{x_i})}{P(Y_i = r|\boldsymbol{x_i})} = \frac{\frac{e^{(x_i'\boldsymbol{\beta_j})}}{\sum_{m=1}^{J} e^{(x_i'\boldsymbol{\beta_m})}}}{\frac{e^{(x_i'\boldsymbol{\beta_r})}}{\sum_{m=1}^{J} e^{(x_i'\boldsymbol{\beta_m})}}} = e^{x_i'(\boldsymbol{\beta_j}-\boldsymbol{\beta_r})} = e^{x_i'(\boldsymbol{\beta_j}-\boldsymbol{\beta_r})}$$

- Para el caso de un modelo *logit* multinomial, se puede ver como la *odds ratio* solo depende de j y r, por lo que las otras alternativas son irrelevantes y los coeficientes estimados para las otras alternativas no tendrían que variar nada o por lo menos no mucho si se extraen alternativas

$$\frac{P\left(Y_i = j\middle|\boldsymbol{x_i}, \boldsymbol{z_{ij}}\right)}{P(Y_i = r|\boldsymbol{x_i}, \boldsymbol{z_{ir}})} = \frac{\frac{e^{\left[x_i'\boldsymbol{\beta_j}+z_{ij}'\boldsymbol{\alpha}\right]}}{\sum_{m=1}^{J} e^{[x_i'\boldsymbol{\beta_m}+z_{im}'\boldsymbol{\alpha}]}}}{\frac{e^{[x_i'\boldsymbol{\beta_r}+z_{ir}'\boldsymbol{\alpha}]}}{\sum_{m=1}^{J} e^{[x_i'\boldsymbol{\beta_m}+z_{im}'\boldsymbol{\alpha}]}}} =$$

$$= \frac{e^{\left[x_i'\boldsymbol{\beta_j}+z_{ij}'\boldsymbol{\alpha}\right]}}{e^{[x_i'\boldsymbol{\beta_r}+z_{ir}'\boldsymbol{\alpha}]}} = \frac{e^{(x_i'\boldsymbol{\beta_j})}e^{(z_{ij}'\boldsymbol{\alpha})}}{e^{(x_i'\boldsymbol{\beta_j})}e^{(z_{ir}'\boldsymbol{\alpha})}} = e^{x_i'(\boldsymbol{\beta_j}-\boldsymbol{\beta_r})+\left(z_{ij}'-z_{ir}'\right)\boldsymbol{\alpha}}$$

- En ambos casos, la variación en la *odds ratio* por el incremento de una unidad de $\boldsymbol{X}$ es la exponencial de la diferencia de los coeficientes $\boldsymbol{\beta}$ de cada alternativa. En el caso del *logit* condicional, un aumento en $\boldsymbol{Z}$ provoca un incremento de $e^{(\boldsymbol{z_{ij}'}-\boldsymbol{z_{ir}'})\boldsymbol{\alpha}}$ en la *odds ratio*

$$\Delta X_{ij} = 1 \Rightarrow \frac{P(Y_i = j|\boldsymbol{x_i})}{P(Y_i = r|\boldsymbol{x_i})} * e^{x_i'(\boldsymbol{\beta_j}-\boldsymbol{\beta_r})} = e^{(x_i'+\boldsymbol{1'})(\boldsymbol{\beta_j}-\boldsymbol{\beta_r})}$$

$$\Delta X_i = 1 \Rightarrow \frac{P\left(Y_i = j\middle|\boldsymbol{x_i}, \boldsymbol{z_{ij}}\right)}{P(Y_i = r|\boldsymbol{x_i}, \boldsymbol{z_{ir}})} * e^{x_i'(\boldsymbol{\beta_j}-\boldsymbol{\beta_r})}$$

$$= e^{(x_i'+\boldsymbol{1})(\boldsymbol{\beta_j}-\boldsymbol{\beta_g})+\left(z_{ij}'-z_{ig}'\right)\boldsymbol{\alpha}}$$

$$\Delta Z_{ij} = 1 \Rightarrow \frac{P\left(Y_i = j\middle|\boldsymbol{x_i}, \boldsymbol{z_{ij}}\right)}{P(Y_i = r|\boldsymbol{x_i}, \boldsymbol{z_{ir}})} * e^{\left(z_{ij}'-z_{ir}'\right)\boldsymbol{\alpha}}$$

$$= e^{x_i'(\boldsymbol{\beta_j}-\boldsymbol{\beta_g})+2\left(z_{ij}'-z_{ir}'\right)\boldsymbol{\alpha}}$$

- Para saber los cambios probabilísticos que se dan cuando se incluye o se excluye una alternativa en un modelo *logit* multinomial, se resuelve el siguiente sistema:

$$\begin{cases} \sum_{j=1}^{J} P(Y_i = j|\boldsymbol{x_{ij}}) = 1 \\ \dfrac{P(Y_i = j|\boldsymbol{x_{ij}})}{P(Y_i = m|\boldsymbol{x_{ij}})} = a\ when\ j \neq m \\ P(Y_i = j|\boldsymbol{x_{ij}})\ \ for\ j = 1,2,\dots,J \end{cases}$$

- Un inconveniente de esta suposición es que a veces da resultados contrarios al sentido común, debido a que la *odds ratio* de las probabilidades no puede variar y hay casos en los que la inclusión o exclusión de una alternativa hace que varíe la probabilidad de las alternativas

- La IIA nace de la suposición inicial de que los errores son independientes y homoscedásticos. Por lo tanto, para comprobar que se cumple la suposición de IIA, se puede usar el test de Hausman o *Hausman test* y así comprobar la consistencia de los estimadores

$$\begin{cases} H_0: the\ models\ are\ equal \\ H_1: the\ models\ are\ not\ equal \end{cases}$$

$$h = (\boldsymbol{\beta_s} - \boldsymbol{\beta_f})'[\boldsymbol{V_s} - \boldsymbol{V_f}]^{-1}(\boldsymbol{\beta_s} - \boldsymbol{\beta_f}) \sim \chi_r^2$$

$$s = estim.on\ the\ restr.subset \quad f = estim.on\ the\ full\ set$$

$$\boldsymbol{V} = asympt.cov.matrix \qquad r = nº\ of\ elim.coefficients$$

- El test se basa en estimar dos modelos: uno completo y otro eliminando las alternativas que se sospecha que no cumplen la IIA. De este modo, la diferencia no está en las restricciones que se imponen en coeficientes (como en otros contrastes de hipótesis), si no en el tamaño de la muestra con el que se trabaja (se eliminan las observaciones para las alternativas sospechosas)

- Si se cumple, los modelos tienen que ser muy parecidos o iguales, dado que la eliminación de las alternativas irrelevantes no tiene que afectar a la probabilidad de escoger entre un par de alternativas (la probabilidad relativa de escoger una alternativa *j*) y los estimadores serán consistentes. La hipótesis nula del test expresa que los dos modelos son iguales o muy parecidos (se cumple la IIA), mientras que la hipótesis alternativa expresa que no

- Los coeficientes que se eliminan serán el número de regresores más una unidad $(k + 1)$ por cada alternativa que se elimina,

debido a que los coeficientes para cada alternativa son diferentes (pero no varía el número de regresores)

$$r = (k+1) * nº\ of\ eliminated\ altern. + d$$

$$d = number\ of\ \boldsymbol{z}\ regressors\ in\ remaining\ altern.$$

- Las medidas de ajuste y los tests de hipótesis utilizadas en el modelo *logit* binomial pueden utilizarse perfectamente para los modelos *logit* multinomial y para los modelos *logit* condicional, pero las restricciones se ven afectadas

    - En la *likelihood ratio test*, por ejemplo, dado que habrá un número k de coeficientes de regresores para cada alternativa, los grados de libertad serán el producto entre el número de alternativas y el número de regresores

$$q = k * nº\ of\ alternatives \rightarrow \quad Multinomial\ Logit$$

$$q = k * nº\ of\ alternatives + d^{tot} \rightarrow \quad Conditional\ Logit$$

$$d^{tot} = number\ of\ \boldsymbol{z}\ regressors$$

- Si no se cumple la condición de IIA, se necesita utilizar una alternativa al modelo *logit* multinomial o condicional. Una manera de relajar la condición es el modelo *logit* anidado, de modo que se agrupan alternativas similares en subgrupos para que se mantenga la IIA en cada uno

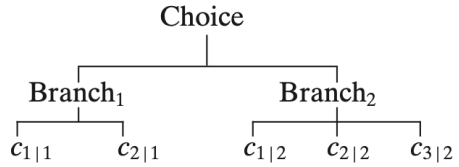
    - Se supone que hay J alternativas y que se pueden dividir en B subgrupos o ramas por similitud, de modo que el conjunto de elección se puede dividir de la siguiente manera:

$$\{c_1, c_2, \dots, c_J\} = \left\{\{c_{1|1}, \dots, c_{J_1|1}\}, \{c_{1|2}, \dots, c_{J_2|2}\}, \dots, \{c_{1|B}, \dots, c_{J_B|B}\}\right\}$$

        - Cada subgrupo tiene un número de alternativas similares J_b que no son las mismas para cada grupo, por lo que el subíndice $j|b$ indica el número de la alternativa dentro del subgrupo, pero no indica el número de la alternativa dentro del conjunto original de alternativas

$$c_{j|b} \neq c_j$$

        - Se puede entender que el proceso de elección consiste en escoger un subconjunto de los B posibles y entonces seleccionar una alternativa dentro de cada subgrupo



- Suponiendo que existen observaciones para los atributos de las alternativas $\boldsymbol{X_{ij|b}}$ y para atributos de los subconjuntos de elección $\boldsymbol{Z_{ib}}$, el modelo econométrico permite obtener la siguiente probabilidad incondicional para la alternativa j_b del subgrupo b

$$P(j_b, b|\boldsymbol{x_{ij|b}}, \boldsymbol{z_{ib}}) = \frac{e^{\left(\boldsymbol{x'_{ij|b}\beta + z'_{ib}\alpha}\right)}}{\sum_{l=1}^{B}\sum_{r=1}^{J_l} e^{\left(\boldsymbol{x'_{ir|l}\beta + z'_{il}\alpha}\right)}}$$

  - En este caso, se asume que los parámetros de los coeficientes son constantes, de modo que no dependen de las alternativas j ni de los subgrupos b, por lo que es parecido a un modelo *logit* condicional

$$U_{ij} = \boldsymbol{x'_{ij|b}\beta + z'_{ib}\alpha} + \varepsilon_{ib}$$

  - El término de error en este modelo sigue una distribución de valor extremo generalizado o *generalized extreme value* (GEV)

$$\varepsilon_{ib} \sim F(\boldsymbol{\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iB}}) \quad where$$

$$F(\boldsymbol{\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iB}}) = \exp\left[\sum_{l=1}^{B}\left(\sum_{r=1}^{J_l} e^{-\frac{\boldsymbol{\varepsilon_{kr}}}{\tau_l}}\right)^{\tau_l}\right] \quad for\ all\ k \in B$$

- Esta probabilidad se puede descomponer en la probabilidad de escoger la alternativa j_b y la probabilidad de escoger el subgrupo b

$$P(j_b, b|\boldsymbol{x_{ij|b}}, \boldsymbol{z_{ib}}) = \frac{e^{\left(\boldsymbol{x'_{ij|b}\beta + z'_{ib}\alpha}\right)}}{\sum_{l=1}^{B}\sum_{r=1}^{J_l} e^{\left(\boldsymbol{x'_{ir|l}\beta + z'_{il}\alpha}\right)}} =$$

$$= \frac{e^{\left(\boldsymbol{x'_{ij|b}\beta}\right)} e^{\left(\boldsymbol{z'_{ib}\alpha}\right)}}{\sum_{l=1}^{B}\sum_{r=1}^{J_l} e^{\left(\boldsymbol{x'_{ir|l}\beta}\right)} e^{\left(\boldsymbol{z'_{il}\alpha}\right)}} = \frac{e^{\left(\boldsymbol{x'_{ij|b}\beta}\right)}}{\sum_{r=1}^{J_b} e^{\left(\boldsymbol{x'_{ir|b}\beta}\right)}} \frac{e^{\left(\boldsymbol{z'_{ib}\alpha}\right)}}{\sum_{l=1}^{B} e^{\left(\boldsymbol{z'_{il}\alpha}\right)}} \frac{\left[\sum_{r=1}^{J_b} e^{\left(\boldsymbol{x'_{ir|b}\beta}\right)}\right]\left[\sum_{l=1}^{B} e^{\left(\boldsymbol{z'_{il}\alpha}\right)}\right]}{\sum_{l=1}^{B}\sum_{r=1}^{J_b} e^{\left(\boldsymbol{x'_{ir|l}\beta}\right)} e^{\left(\boldsymbol{z'_{il}\alpha}\right)}}$$

$$= P\left[Y_i = j_b|\boldsymbol{x_{ij|b}}, \boldsymbol{z_{ib}}, b\right] P\left[b|\boldsymbol{x_{ij|b}}, \boldsymbol{z_{ib}}\right]$$

$$where \begin{cases} P\left(Y_i = j_b \middle| \boldsymbol{x_{ij|b}}, \boldsymbol{z_{ib}}, b\right) = \dfrac{e^{\left(\boldsymbol{x'_{ij|b}\beta}\right)}}{\sum_{r=1}^{J_b} e^{\left(\boldsymbol{x'_{ir|b}\beta}\right)}} \\ P\left(b \middle| \boldsymbol{x_{ij|b}}, \boldsymbol{z_{ib}}\right) = \dfrac{e^{\left(z'_{ib}\alpha\right)}}{\sum_{l=1}^{B} e^{\left(z'_{il}\alpha\right)}} \dfrac{\left[\sum_{r=1}^{J_b} e^{\left(x'_{ir|b}\beta\right)}\right]\left[\sum_{l=1}^{B} e^{\left(z'_{il}\alpha\right)}\right]}{\sum_{l=1}^{B}\sum_{r=1}^{J_b} e^{\left(x'_{ir|l}\beta\right)} e^{\left(z'_{il}\alpha\right)}} \end{cases}$$

- La variable IV_{ib} se denomina valor inclusivo o *inclusive value*, y estos dependen del individuo y del subgrupo b, pero no de la alternativa (por lo que se tiene que calcular para cada alternativa)

$$IV_{ib} = \ln\left(\sum_{r=1}^{J_b} e^{\left(\boldsymbol{x'_{ir|b}\beta}\right)}\right)$$

$$\Rightarrow P\left(j_b, b \middle| \boldsymbol{x_{ij|b}}, \boldsymbol{z_{ib}}\right) = \frac{e^{\left(\boldsymbol{x'_{ij|b}\beta}\right)}}{\sum_{r=1}^{J_b} e^{\left(\boldsymbol{x'_{ir|b}\beta}\right)}} \frac{e^{\left[\tau_b\left(\boldsymbol{z'_{ib}\alpha + IV_{ib}}\right)\right]}}{\sum_{l=1}^{B} e^{\left[\tau_l\left(\boldsymbol{z'_{il}\alpha + IV_{il}}\right)\right]}}$$

$$\Rightarrow \begin{cases} P\left(Y_i = j_b \middle| \boldsymbol{x_{ij|b}}, \boldsymbol{z_{ib}}, b\right) = \dfrac{e^{\left(\boldsymbol{x'_{ij|b}\beta}\right)}}{\sum_{r=1}^{J_b} e^{\left(\boldsymbol{x'_{ir|b}\beta}\right)}} \\ P\left(b \middle| \boldsymbol{x_{ij|b}}, \boldsymbol{z_{ib}}\right) = \dfrac{e^{\left[\tau_b\left(\boldsymbol{z'_{ib}\alpha + IV_{ib}}\right)\right]}}{\sum_{l=1}^{B} e^{\left[\tau_l\left(\boldsymbol{z'_{il}\alpha + IV_{il}}\right)\right]}} \end{cases}$$

- Para que el modelo derivado con el valor inclusivo sea igual al modelo para la probabilidad incondicional, los valores τ deben ser iguales a 1. Por lo tanto, se utiliza una restricción $\tau = 1$ para que el modelo sea equivalente al de un *logit* condicional

- No obstante, si los valores τ no son unitarios, entonces nace un modelo diferente, que es el modelo *logit* anidado. Estos valores τ son el grado de disimilitud, siendo una medida inversa de la correlación entre los errores de diferentes alternativas en un mismo grupo

$$\tau_b = \sqrt{1 - \rho_{j,k}} \quad where \;\; \rho_{j,k} = corr\left(\varepsilon_{bj}, \varepsilon_{bk}\right) \; and \; j \neq k$$

- La correlación entre los errores de un mismo grupo tiene que tener correlación positiva, dado que las alternativas dentro de cada grupo son similares. Por lo tanto, τ_b tiene que estar necesariamente entre 0 y 1

- Para grupos que tengan una sola alternativa, la correlación no estará definida (porque no hay otras alternativas) y por tanto se asume que $\tau = 1$ para estos grupos

- Cuando hay un subgrupo formado por subgrupos, se utilizan varios valores τ_{hb}, donde h indica el subgrupo superior (formado por subgrupos inferiores) y b indica el subgrupo inferior

- Debido a que la IIA se debe a la independencia y homoscedasticidad de los errores, al agrupar alternativas en subgrupos se permite que la varianza difiera entre cada subgrupo pero que se mantenga la homoscedasticidad dentro de cada uno

  - Si los valores τ difieren de 1 (se relaja la restricción), el modelo es un *logit* anidado. Los parámetros τ permiten que exista heterogeneidad entre grupos, pero homoscedasticidad dentro de cada uno, dado que la varianza dentro de cada grupo se define con la siguiente expresión:

$$\sigma_b^2 = \frac{\pi^2}{6\tau_b}$$

- Como los coeficientes no son interpretables, se utilizan los efectos marginales. Comúnmente, se utiliza la semielasticidad para poder entender el efecto de un aumento de un incremento unitario en las variables $\boldsymbol{X_{ij|b}}$

$$\frac{\partial \ln\left[P\left(j_b, b|\boldsymbol{x_{ij|b}}, \boldsymbol{z_{ib}}\right)\right]}{\partial \boldsymbol{x_{ij|b}}} =$$

$$= \left\{\mathbf{1}(l = b)\left[\mathbf{1}(r = j_b) - P_{j_b|b}\right] + \tau_b[\mathbf{1}(l = b) - P_b]P_{j_b|b}\right\}\boldsymbol{\beta}$$

$$where\ \ P_{j_b|b} = P\left(Y_i = j_b \middle| \boldsymbol{x_{ij|b}}, \boldsymbol{z_{ib}}, b\right)\ \ and\ \ P_b = P\left(b \middle| \boldsymbol{x_{ij|b}}, \boldsymbol{z_{ib}}\right)$$

  - Si el grupo no es el mismo, entonces el efecto tiene el signo opuesto a $\boldsymbol{\beta}$, petro

- Para poder estimar los coeficientes de un modelo *logit* anidado se pueden estimar de dos maneras diferentes:

  - Un enfoque de máxima verosimilitud de dos pasos (llamado enfoque de información limitada) que consiste en estimar las $\boldsymbol{\beta}$ tratando la elección dentro de cada subgrupo como un modelo *logit* condicional, y después calcular los valores intrínsecos para cada subgrupo y estimar las $\boldsymbol{\alpha}$ y $\boldsymbol{\tau}$ tratando la elección entre grupos como un modelo *logit* condicional con atributos $\boldsymbol{z_{ib}}$ e $\boldsymbol{IV_{ib}}$

- Un enfoque de información completa de máxima verosimilitud, en donde se maximiza la función logarítmica de máxima verosimilitud

$$\ln L = \sum_{i=1}^{n} \ln\left[P\left(Y_i = j_b \middle| \boldsymbol{x_{ij|b}}, \boldsymbol{z_{ib}}, b\right) P\left(b \middle| \boldsymbol{x_{ij|b}}, \boldsymbol{z_{ib}}\right)\right]$$

- El enfoque de información limitada es menos eficiente debido a que la estimación por dos pasos hace que la matriz de estimaciones se tenga que corregir en el segundo paso

  - Para poder contrastar hipótesis sobre si el modelo tiene que ser anidado o no, se tienen que hacer tests de hipótesis para la heteroscedasticidad, la cual se puede comprobar a través de tres estadísticos diferentes

    - El test de hipótesis de Wald es un test de heteroscedasticidad que utiliza

- Otra variante del modelo de elección multinomial es el modelo *logit* de parámetros aleatorios o *random parameters logit model*, también llamado modelo mixto

  - En una posible especificación, se considera que los coeficientes son aleatorios (por lo que tienen una formulación propia) y que siguen una distribución normal, de modo que hay un modelo para la utilidad y otro modelo para los coeficientes

$$U_{ij} = \boldsymbol{z'_{ij}\beta_i} + \varepsilon_{ij} \quad where \quad \beta_i = \beta + u_i$$

$$u_i \sim N(0, \sigma_u^2)$$

    - Dado que los coeficientes son aleatorios y dependen de un término de error distribuido normalmente, los coeficientes aleatorios también seguirán una distribución normal con media β y matriz de covarianzas $\boldsymbol{\Sigma_\beta}$

$$\beta_i \sim N\left(\beta, \boldsymbol{\Sigma_\beta}\right)$$

  - Otra especificación asume que los coeficientes son constantes pero el término de error está determinado por las mismas variables $\boldsymbol{Z_{ij}}$ y un término de error aleatorio

$$U_{ij} = \boldsymbol{z'_{ij}\beta} + \eta_{ij} \quad where \quad \eta_{ij} = \boldsymbol{z'_{ij}\gamma_i} + \varepsilon_{ij}$$

- Como U_{ij} y η_{ij} están determinadas por las mismas variables, ambas variables están correlacionadas y su correlación se puede expresar de la siguiente manera:

$$Cov(\eta_{ij}, \eta_{ig}) = \boldsymbol{z'_{ij}\Sigma_{\beta}z_{ij}}$$

- Finalmente, una especificación más general es una en la que se tiene un modelo para la utilidad, que depende de las variables $\boldsymbol{x_{ij}}$ y de los parámetros $\boldsymbol{\beta_{ik}}$, los cuales se determinan por unas variables $\boldsymbol{z_i}$ y de los parámetros $\boldsymbol{\theta}$

$$U_{ij} = \boldsymbol{x'_{ij}\beta_i} + \varepsilon_{ij} \quad where \quad \beta_i = \beta + \boldsymbol{z'_i\theta} + u_{ij}$$

$$u_i \sim N(0, \sigma_u^2)$$

  - La distribución de los parámetros $\boldsymbol{\beta_i}$ será una distribución normal con los siguientes estadísticos:

$$E(\boldsymbol{\beta_i}) = \beta + E(\boldsymbol{z'_i})\boldsymbol{\theta} \qquad Var(\boldsymbol{\beta_i}) = Var(\boldsymbol{z'_i})\boldsymbol{\theta^2}$$

- La probabilidad de que se escoja una alternativa *j* se puede calcular a través de la esperanza condicional en las variables explicativas y $\boldsymbol{u_i}$

$$E\left[P\left(Y_i = j|\boldsymbol{x_{ij}}, \boldsymbol{z_i}, \boldsymbol{u_i}\right)\right] = \int_1^N \frac{e^{\left(\boldsymbol{x'_{ij}\beta_i}\right)}}{\sum_{l=1}^{J} e^{\left(\boldsymbol{x'_{il}\beta_i}\right)}}\, \phi(\boldsymbol{\beta_i})\, d\boldsymbol{\beta_i}$$

$$where \;\; \boldsymbol{u_i} = \begin{pmatrix} u_{i1} \\ \cdots \\ u_{iJ} \end{pmatrix}$$

  - Debido a que los coeficientes dependen del término de error u_{ij}, se tiene que calcular la esperanza condicional a estos con tal de que las $\boldsymbol{\beta_i}$ sean constantes para una u_{ij} determinada

  - Para poder estimar esta probabilidad, se puede utilizar métodos numéricos. De este modo, se pueden utilizar las estimaciones para poder estimar la función logarítmica de máxima verosimilitud

$$\hat{E}\left[P\left(Y_i = j|\boldsymbol{x_{ij}}, \boldsymbol{z_i}, \boldsymbol{u_i}\right)\right] = \frac{1}{S}\sum_{s=1}^{S} \frac{e^{\left(\boldsymbol{x'_{ij}\beta_i^{(s)}}\right)}}{\sum_{l=1}^{J} e^{\left(\boldsymbol{x'_{il}\beta_i^{(s)}}\right)}}$$

$$\ln \hat{L} = \sum_{i=1}^{N}\sum_{j=1}^{J} y_{ij} \ln\left\{\hat{E}\left[P\left(Y_i = j|\boldsymbol{x_{ij}}, \boldsymbol{z_i}, \boldsymbol{u_i}\right)\right]\right\}$$

Los modelos de elección discreta multinomial ordenada

- Hay veces que las alternativas de la variable dependiente discreta están ordenadas, por lo que se revela la fuerza de las preferencias con respecto a un único resultado y se puede explotar este aspecto con modelos de elección discreta multinomial ordenada

    - Para cualquier individuo, se supone que hay una variable de utilidad continua subyacente que revela la fuerza de la preferencia y que depende de variables independientes para cada individuo

$$U_{ij}^{*} = \boldsymbol{x_i'\beta} + \varepsilon_{ij} \quad where \quad \boldsymbol{x_i} = \begin{pmatrix} 1 \\ X_{i1} \\ \dots \\ X_{iK} \end{pmatrix} \; and \; \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_K \end{pmatrix}$$

        - El modelo sugiere que cada individuo tiene unas características que determinan su utilidad para la alternativa j pero solo dependen del individuo, y un término de error que depende del individuo y de la alternativa j

        - Se asume que la variable subyacente o de utilidad se distribuye por toda la recta real

$$-\infty < U_{ij}^{*} < \infty$$

    - La existencia de un orden entre las alternativas se traduce en la censura de esta variable subyacente, de modo que la elección de j es una versión censurada de la utilidad

$$Y_i = j \quad if \quad \mu_{j-1} < U_{ij}^{*} \leq \mu_j \quad for \; j = 1, \dots, J$$

$$when \; \mu_0 = -\infty \quad and \quad \mu_J = \infty$$

        - De este modo, cada alternativa j se escoge cuando la utilidad subyacente está dentro del intervalo de esa alternativa

$$U_{ij} \in (\mu_{j-1}, \mu_j]$$

        - Existirán $J - 1$ límites μ para dividir los intervalos de la variable subyacente para los cuales se identifican las observaciones

    - La diferencia entre dos alternativas en la jerarquía o estructura de la variable dependiente (una alternativa j comparada con una alternativa g) no es la misma que en la escala de la utilidad subyacente

- Eso quiere decir que se ha hecho una transformación estrictamente no lineal capturada por los límites μ, por lo que se pueden estimar los coeficientes con un modelo de elección ordenada

- El modelo *probit* ordenado se construye en base a una regresión latente similar a la del modelo *probit* binomial, solo que en este se tiene en cuenta que las alternativas están ordenadas en la variable dependiente

  - Se asume que existe una variable no observada U_{ij}^* para un individuo i y una alternativa j, la cual depende de características individuales y de un término de error que también depende de la alternativa j

$$U_{ij}^* = \boldsymbol{x_i'\beta} + \varepsilon_{ij} \quad where \quad \boldsymbol{x_i} = \begin{pmatrix} 1 \\ X_{i1} \\ \dots \\ X_{iK} \end{pmatrix} \; and \; \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_K \end{pmatrix}$$

    - Se asume que el término de error se distribuye de manera normal con media nula y varianza unitaria

$$\varepsilon_{ij} \sim N(0,1)$$

    - No obstante, se observa una variable Y_i, la cual está censurada para diferentes intervalos que están determinados por los límites μ, los cuales son desconocidos y se tienen que estimar con las $\boldsymbol{\beta}$

$$Y_i = j \quad if \quad \mu_{j-1} < U_{ij}^* \leq \mu_j \quad for \; j = 1, \dots, J$$

$$when \quad \mu_0 = -\infty \quad and \; \mu_J = \infty$$

  - A partir de las suposiciones, se puede ver que la probabilidad de que se escoja una alternativa j será la probabilidad de que el término de error esté dentro de un intervalo determinado por los límites μ y las características individuales $\boldsymbol{X_i}$

$$P(Y_i = j|\boldsymbol{x_i}) = P\left(\mu_{j-1} < U_{ij}^* \leq \mu_j \mid \boldsymbol{x_i}\right) = P\left(\mu_{j-1} < \boldsymbol{x_i'\beta} + \varepsilon_{ij} \leq \mu_j \mid \boldsymbol{x_i}\right)$$

$$= P\left(\mu_{j-1} - \boldsymbol{x_i'\beta} < \varepsilon_{ij} \leq \mu_j - \boldsymbol{x_i'\beta} \mid \boldsymbol{x_i}\right) = \Phi\left(\mu_j - \boldsymbol{x_i'\beta}\right) - \Phi\left(\mu_{j-1} - \boldsymbol{x_i'\beta}\right)$$

    - Como las probabilidades tienen que ser positivas, es necesario que los límites sean estrictamente crecientes

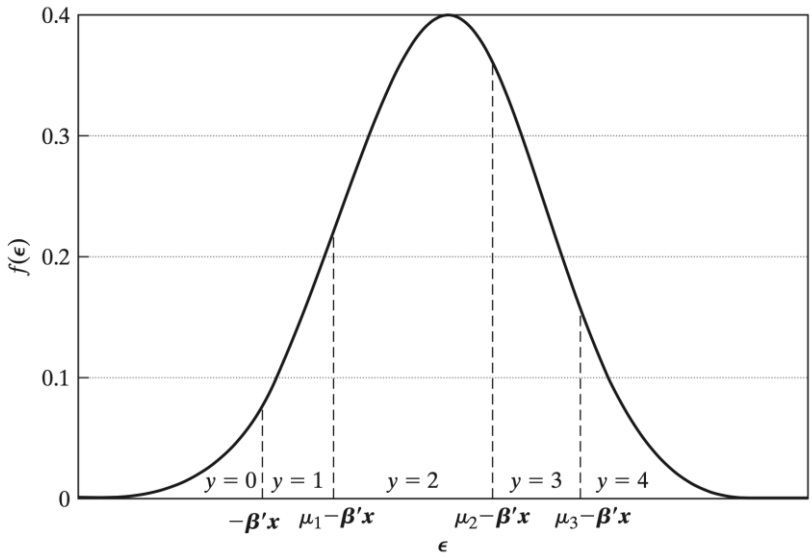
$$\mu_1 < \mu_2 < \dots < \mu_{J-1} < \mu_J$$

- Debido a que se asume que $\mu_0 = -\infty$ y que $\mu_J = \infty$, se puede ver como la probabilidad de la primera opción y de la última se pueden expresar de la siguiente manera:

$$P(Y_i = 1|\boldsymbol{x_i}) = \Phi(\mu_1 - \boldsymbol{x_i'\beta})$$

$$P(Y_i = J|\boldsymbol{x_i}) = 1 - \Phi\big(\mu_{J-1} - \boldsymbol{x_i'\beta}\big)$$

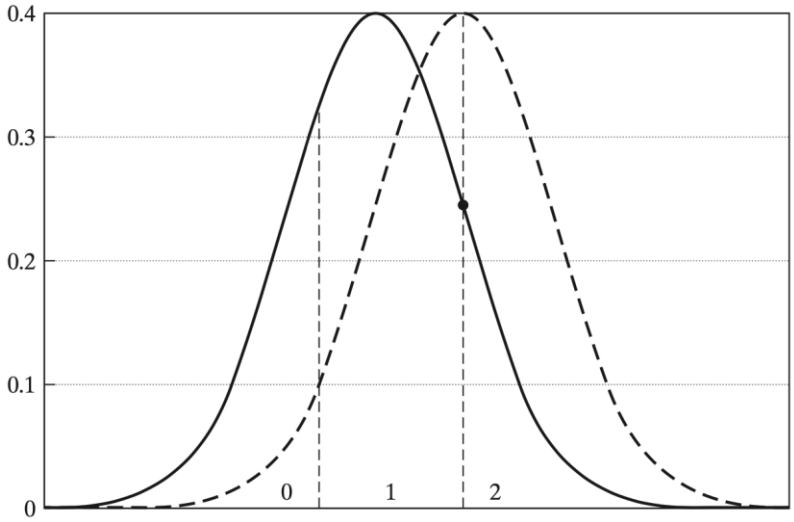
- La estructura del modelo permite ver como la distribución del error permite obtener la probabilidad para una alternativa, en donde el valor z depende de μ y de $\boldsymbol{x_i'\beta}$



- Los efectos marginales en el modelo dependerán de μ y de $\boldsymbol{x_i'\beta}$, dado que la probabilidad depende de la función de distribución para $\mu - \boldsymbol{x_i'\beta}$, y eso hace que se tenga una interpretación diferente

$$\frac{\partial P(Y_i = j|\boldsymbol{x_i})}{\partial X_{ik}} = \big[\phi\big(\mu_{j-1} - \boldsymbol{x_i'\beta}\big) - \phi\big(\mu_j - \boldsymbol{x_i'\beta}\big)\big]\beta_{ik}$$

  - Siendo μ y $\boldsymbol{\beta}$ constantes, un incremento unitario de una variable X_{ik} es equivalente a desplazar la distribución, por lo se reduce la masa de probabilidad de la primera o la última categoría (dependiendo de $\boldsymbol{\beta}$) y eso hace que aumenten y disminuyan las probabilidades de diferentes categorías



- Debido a que el efecto marginal de la primera categoría tiene un signo contrario al del coeficiente β_{ik}, mientras que el de la última tiene el mismo signo, el desplazamiento de la masa provocado por un incremento unitario en X_{ik} hace que se tenga un efecto opuesto en la probabilidad entre la primera y la última categoría

$$If\ \beta_{ik} > 0 \ \Rightarrow \begin{cases} \Delta P(Y_i = 1|\boldsymbol{x_i}) < 0 \\ \Delta P(Y_i = J|\boldsymbol{x_i}) > 0 \end{cases}$$

$$If\ \beta_{ik} < 0 \ \Rightarrow \begin{cases} \Delta P(Y_i = 1|\boldsymbol{x_i}) > 0 \\ \Delta P(Y_i = J|\boldsymbol{x_i}) < 0 \end{cases}$$

$$\frac{\partial P(Y_i = 1|\boldsymbol{x_i})}{\partial X_{ik}} = -\phi\left(\mu_j - \boldsymbol{x_i'\beta}\right)\beta_{ik}$$

$$\frac{\partial P(Y_i = J|\boldsymbol{x_i})}{\partial X_{ik}} = \phi\left(\mu_{j-1} - \boldsymbol{x_i'\beta}\right)\beta_{ik}$$

- No obstante, el efecto en las categorías intermedias depende de dos valores diferentes ambas funciones de densidad, de modo que el efecto es ambiguo si no se estima

- El modelo *logit* ordenado se construye en base a una regresión latente similar a la del modelo *logit* binomial, solo que en este se tiene en cuenta que las alternativas están ordenadas en la variable dependiente

  - Se asume que existe una variable no observada U_{ij}^* para un individuo i y una alternativa j, la cual depende de características individuales y de un término de error que también depende de la alternativa j

$$U_{ij}^{*} = \boldsymbol{x_i'\beta} + \varepsilon_{ij} \quad where \quad \boldsymbol{x_i} = \begin{pmatrix} 1 \\ X_{i1} \\ \dots \\ X_{iK} \end{pmatrix} \; and \; \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_K \end{pmatrix}$$

- Se asume que el término de error se distribuye como en un modelo *logit* multinomial

$$\varepsilon_{ij} \sim Gumbel \qquad F(\varepsilon_{ij}) = \exp\left(-\frac{1}{e^{\varepsilon_{ij}}}\right)$$

- No obstante, se observa una variable Y_i, la cual está censurada para diferentes intervalos que están determinados por los límites μ, los cuales son desconocidos y se tienen que estimar con las $\boldsymbol{\beta}$

$$Y_i = j \quad if \quad \mu_{j-1} < U_{ij}^{*} \leq \mu_j \quad for \; j = 1, \dots, J$$

$$when \quad \mu_0 = -\infty \quad and \quad \mu_J = \infty$$

- A partir de las suposiciones, se puede ver que la probabilidad de que se escoja una alternativa *j* será la probabilidad de que el término de error esté dentro de un intervalo determinado por los límites μ y las características individuales $\boldsymbol{X_i}$

$$P(Y_i = j|\boldsymbol{X_i}) = P\left(\mu_{j-1} < U_{ij}^{*} \leq \mu_j \,|\boldsymbol{x_i}\right) = P\left(\mu_{j-1} < \boldsymbol{x_i'\beta} + \varepsilon_{ij} \leq \mu_j \mid \boldsymbol{x_i}\right)$$

$$= P\left(\mu_{j-1} - \boldsymbol{x_i'\beta} < \varepsilon_{ij} \leq \mu_j - \boldsymbol{x_i'\beta} \mid \boldsymbol{x_i}\right) = \frac{e^{(\mu_j - \boldsymbol{x_i'\beta})} - e^{(\mu_{j-1} - \boldsymbol{x_i'\beta})}}{\sum_{r=1}^{J} e^{(\mu_r - \boldsymbol{x_i'\beta})}}$$

- Como las probabilidades tienen que ser positivas, es necesario que los límites sean estrictamente crecientes

$$\mu_1 < \mu_2 < \dots < \mu_{J-1} < \mu_J$$

- Debido a que se asume que $\mu_0 = -\infty$ y que $\mu_J = \infty$, se puede ver como la probabilidad de la primera opción y de la última se pueden expresar de la siguiente manera:

$$P(Y_i = 1|\boldsymbol{X_i}) = \frac{e^{(\mu_1 - \boldsymbol{x_i'\beta})}}{\sum_{r=1}^{J} e^{(\mu_r - \boldsymbol{x_i'\beta})}}$$

$$P(Y_i = J|\boldsymbol{X_i}) = 1 - \frac{e^{(\mu_{J-1} - \boldsymbol{x_i'\beta})}}{\sum_{r=1}^{J} e^{(\mu_r - \boldsymbol{x_i'\beta})}}$$

- La estructura del modelo permite ver como la distribución del error permite obtener la probabilidad para una alternativa, en donde el valor z depende de μ y de $\boldsymbol{x_i'\beta}$

- Los efectos marginales en el modelo dependerán de μ y de $\boldsymbol{x_i'\beta}$, dado que la probabilidad depende de la función de distribución para $\mu - \boldsymbol{x_i'\beta}$, y eso hace que se tenga una interpretación diferente

$$\frac{\partial P(Y_i = j|\boldsymbol{x_i})}{\partial X_{ik}} = \left[f\left(\mu_{j-1} - \boldsymbol{x_i'\beta}\right) - f\left(\mu_j - \boldsymbol{x_i'\beta}\right)\right]\beta_{ik}$$

  - Debido a que el efecto marginal de la primera categoría tiene un signo contrario al del coeficiente β_{ik}, mientras que el de la última tiene el mismo signo, el desplazamiento de la masa provocado por un incremento unitario en X_{ik} hace que se tenga un efecto opuesto en la probabilidad entre la primera y la última categoría

$$If\ \beta_{ik} > 0 \ \Rightarrow \begin{cases} \Delta P(Y_i = 1|\boldsymbol{x_i}) < 0 \\ \Delta P(Y_i = J|\boldsymbol{x_i}) > 0 \end{cases}$$

$$If\ \beta_{ik} < 0 \ \Rightarrow \begin{cases} \Delta P(Y_i = 1|\boldsymbol{x_i}) > 0 \\ \Delta P(Y_i = J|\boldsymbol{x_i}) < 0 \end{cases}$$

$$\frac{\partial P(Y_i = 1|\boldsymbol{x_i})}{\partial X_{ik}} = -f\left(\mu_j - \boldsymbol{x_i'\beta}\right)\beta_{ik}$$

$$\frac{\partial P(Y_i = 1|\boldsymbol{x_i})}{\partial X_{ik}} = f\left(\mu_{j-1} - \boldsymbol{x_i'\beta}\right)\beta_{ik}$$

  - No obstante, el efecto en las categorías intermedias depende de dos valores diferentes ambas funciones de densidad, de modo que el efecto es ambiguo si no se estima

- La formulación del modelo de elección ordenada tiene una suposición implícita debido a como se construye, llamada suposición de regresiones paralelas o *parallel regression assumption*

  - A partir del modelo, es posible construir las siguientes variables binarias:

$$w_{ij} = \begin{cases} 1 \ \ if \ \ Y_i > j \\ 0 \ \ if \ \ Y_i \leq j \end{cases} \quad for \ \ \ j = 1,2, \dots, J-1$$

$$with \ \ \ P\left(w_{ij} = 1|\boldsymbol{x_i}\right) = F\left(\boldsymbol{x_i'\beta} - \mu_j\right)$$

  - En este caso, la función de distribución de probabilidad depende del modelo utilizado

$$F(\boldsymbol{x_i'\beta} - \mu_j) = \begin{cases} \Phi(\boldsymbol{x_i'\beta} - \mu_j) \;\; if\; ord.probit \\ \frac{e^{(x_i'\boldsymbol{\beta} - \mu_j)}}{\sum_{r=1}^{J} e^{(x_i'\boldsymbol{\beta} - \mu_r)}} \;\; if\; ord.logit \end{cases}$$

- Se asume que la probabilidad de que $w_{ij} = 1$ es la probabilidad de que $Y_i > j$ para $j = 1,2, \dots, J-1$

$$P(Y_i \leq j|\boldsymbol{x_i}) = P(w_{ij} = 0|\boldsymbol{x_i}) = F(\mu_j - \boldsymbol{x_i'\beta})$$

$$P(Y_i > j|\boldsymbol{X_i}) = P(w_{ij} = 1|\boldsymbol{X_i}) = 1 - F(\mu_j - \boldsymbol{x_i'\beta}) = F(\boldsymbol{x_i'\beta} - \mu_j)$$

- Este enfoque muestra como, a partir de w_{ij} y $\boldsymbol{x_i}$, es posible estimar el mismo vector $\boldsymbol{\beta}$ para cada regresión binaria que se haga de w_{ij} sobre $\boldsymbol{x_i}$, aunque el parámetro constante sea diferente para cada una

  - Esta igualdad entre vectores de parámetros es la suposición de regresiones paralelas, en donde el nombre viene del hecho de que la pendiente (el vector) es el mismo para cada regresión, pero el término constante difiere (por lo que las regresiones serían paralelas)

  - Brant, en su investigación de 1990, mostró el caso para el *proportional odds model* para un modelo *logit* ordenado, propuesto por McCullagh en 1980, en donde se obtiene el mismo resultado de manera más clara

$$\frac{P(Y_i \leq j|\boldsymbol{x_i})}{1 - P(Y_i \leq j|\boldsymbol{x_i})} = \frac{P(Y_i \leq j|\boldsymbol{x_i})}{P(Y_i > j|\boldsymbol{x_i})} = e^{(x_i'\boldsymbol{\beta} - \mu_j)}$$

$$\Rightarrow \ln\left[\frac{P(Y_i \leq j|\boldsymbol{x_i})}{P(Y_i > j|\boldsymbol{x_i})}\right] = \boldsymbol{x_i'\beta} - \mu_j$$

- Debido a este hecho, Brant propuso un test de contraste de hipótesis para contrastar si se cumplía la suposición de regresiones paralelas a través de un test de Wald

$$\begin{cases} H_0: \boldsymbol{\beta_1} = \boldsymbol{\beta_2} = \cdots = \boldsymbol{\beta_{J-1}} \\ H_1: Otherwise \end{cases}$$

$$W = (\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}})\left[\boldsymbol{Var}(\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}|\boldsymbol{x_i})\right]^{-1}(\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}})' \sim \chi^2_{(J-2)K}$$

$$where\ \widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}} \\ \dots \\ \widehat{\boldsymbol{\beta}}_{\boldsymbol{J-1}} \end{pmatrix}\ and\ \overline{\boldsymbol{\beta}} = \begin{pmatrix} \boldsymbol{\beta_1} \\ \dots \\ \boldsymbol{\beta_{J-1}} \end{pmatrix}$$

        - En este test, rechazar la hipótesis nula quiere decir que el modelo ordenado no es adecuado para poder modelar los datos, dado que la suposición en la que se basa no se cumple

        - No obstante, el uso de otros modelos en donde $\boldsymbol{\beta}$ dependa de la alternativa j comporta problemas, dado que el modelo latente dependería de j (que se realiza con la realización de Y_i) pero Y_i depende del modelo latente. Esto haría que no se preserve el orden de las probabilidades ni se mantenga la coherencia en el modelo

        - Algunas razones propuestas por las cuales se puede haber rechazado la hipótesis nula son una mala especificación de $\boldsymbol{X_i'\beta}$, la heteroscedasticidad de ε o una mala especificación de la forma distribucional de la variable latente

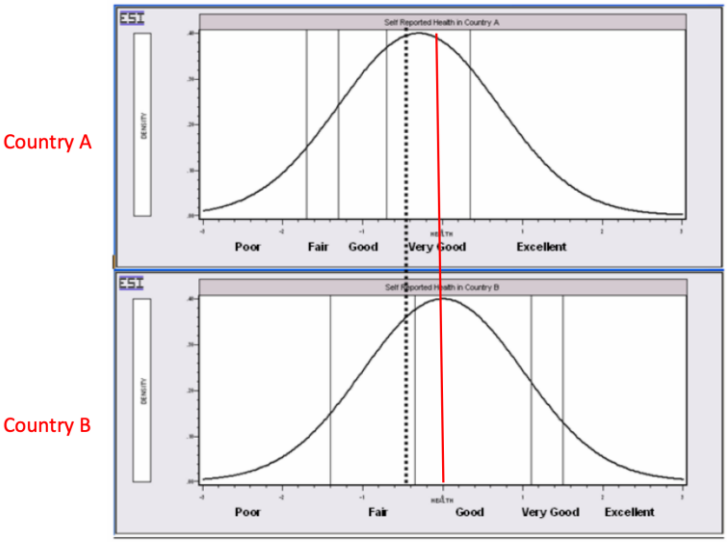
- El modelo estándar para alternativas ordenadas tiene algunas limitaciones a la hora de comparar y de interpretar los resultados

    - Una de las limitaciones más importantes es que hacer comparaciones entre poblaciones o subgrupos de una población puede comportar problemas debidos a sesgos y heterogeneidad

        - La elección de una alternativa puede estar determinada por las alternativas escogidas en el pasado (por ejemplo, con los estados de salud previos). A esto se le llama *dependent-state bias*

        - Dependiendo de la escala de referencia que se utilice, dos individuos que realmente escogerían una misma alternativa pueden escoger dos diferentes. A esto se le llama sesgo de referencia de escala o *scale reference bias*

        - Puede ser que los individuos entiendan las escalas de manera diferente, de modo que dos individuos que realmente escogerían una misma alternativa pueden escoger dos alternativas diferentes debido a su entendimiento. A esto se le llama heterogeneidad de informe o *reporting heterogeneity*



- Otro problema es que diferentes subgrupos de una población utilizan niveles de límites sistemáticamente diferentes para las categorías, aunque realmente pertenezcan a la misma categoría

  - Esto se debe a que cada persona puede interpretar su estado o la categoría en la que está diferente dependiendo de sus características personales

  - Por ello, una solución puede ser incluir la posibilidad de heterogeneidad en los límites, de modo que se definiría el nivel límite de la alternativa j como una función g de las características personales Z_i

$$\mu_{ij} = g(Z_i)$$

Los modelos para el conteo de eventos

- Que la variable dependiente represente el número de ocurrencias de un evento hace que esta sea una medida cuantitativa como las que se utilizan en los modelos de regresión clásicos

  - No obstante, la preponderancia típica de ceros y valores pequeños y la naturaleza discreta de la variable de respuesta hace que se tenga que utilizar un enfoque que tenga en cuenta estos aspectos

    - Uno de los modelos más importantes es el modelo de Poisson, aunque las suposiciones de este han lelvado a que se desarrollen modelos que puedan superar esas limitaciones como el modelo de binomial negativa

- Además, para tener en cuenta la preponderancia de ceros en la variable de respuesta se han desarrollado modelos con datos censurados y de otros tipos

- El modelo de regresión de Poisson especifica que cada y_i proviene de una distribución de Poisson con parámetro λ_i, la cual está relacionada con los regresores $\boldsymbol{x_i}$. La ecuación primordial del modelo es la siguiente:

$$P(Y = y_i|\boldsymbol{x}_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_I!} \quad for\ \ y_i = 0,1,2,\ldots$$

  - La formulación más común para λ_i es el modelo de regresión log-lineal

$$\ln \lambda_i = \boldsymbol{x}_i\boldsymbol{\beta}$$

  - Se puede demostrar fácilmente que el número de eventos esperados por periodos o por unidad de espacio y, por tanto, su varianza condicional (la cual será heterocedástica) y el efecto marginal son los siguientes:

$$E(y_i|\boldsymbol{x}_i) = Var(y_i|\boldsymbol{x}_i) = \lambda_i = e^{\boldsymbol{x}_i'\boldsymbol{\beta}} \qquad \frac{\partial E(y_i|\boldsymbol{x}_i)}{\partial \boldsymbol{x}_i} = \lambda_i\boldsymbol{\beta}$$

  - Si se quisiera obtener el efecto medio del tratamiento, es posible realizar una pequeña modificación al modelo. De este modo, se obtienen las siguientes equivalencias:

$$E(y_i|\boldsymbol{x}_i) = e^{\boldsymbol{x}_i'\boldsymbol{\beta}+\gamma T} \ \Rightarrow\ ATE = \frac{1}{n}\sum_{i=1}^{n}\left[e^{\boldsymbol{x}_i'\boldsymbol{\beta}+\gamma T} - e^{\boldsymbol{x}_i'\boldsymbol{\beta}}\right]$$

  - El ATET se puede calcular a través de hacer esta media con las observaciones en las que $T = 1$

- Debido a que el modelo de Poisson no es lineal, los coeficientes de los regresores no son directamente interpretables como el efecto marginal:

$$\frac{\partial E(y_i|\boldsymbol{x}_i)}{\partial x_k} = \frac{\partial \lambda_i}{\partial x_k} = \beta_k\lambda_i = \beta_k e^{\boldsymbol{x}_i'\boldsymbol{\beta}}$$

  - No se puede interpretar su magnitud (porque no es el coeficiente únicamente), pero si se puede interpretar el signo de los coeficientes

  - El efecto marginal medio estimado, al ser la media de los efectos marginales estimados para cada individuo, también se puede

expresar como el producto entre la media de la variable y y el coeficiente del regresor x_k para el que se estima el efecto

- Aunque la magnitud no sea interpretable en el efecto marginal, sí tiene una interpretación para el efecto sobre el valor esperado de λ_i (dado que depende de los coeficientes)

$$e^{\beta_0+\beta_1 x_1+\cdots+\beta_k(X_k+1)} = e^{\beta_0+\beta_1 X_{1i}+\cdots+\beta_k X_{ki}} e^{\beta_k} = \lambda_i e^{\beta_k}$$

  - De este modo, la variación de una unidad en un regresor hace que $E(y_i|\boldsymbol{x}_i)$ se multiplique por la exponencial del coeficiente del regresor

  - Cuando la variable está en logaritmos, sin embargo, el coeficiente β_k se puede interpretar como la elasticidad gracias a las propiedades de los exponentes

$$y = e^{\ln x_k} \Rightarrow \ln y = \ln x_k$$

- El modelo de Poisson no produce una contraparte natural al R^2 en un modelo de regresión lineal, dado que la media condicional no es lineal y la regresión es heteroscedástica. Por lo tanto, se han propuesto medidas de bondad del ajuste alternativas

  - Se puede utilizar una medida de basada en los residuos estandarizados, la cual compara el ajuste del modelo con el ajuste proporcionado por un modelo solo con el término constante. No obstante, esta medida puede ser negativa y puede aumentar cuando se sacan regresores del modelo

$$R_p^2 = 1 - \frac{\sum_{i=1}^{n}\left[(y_i - \hat{\lambda}_i)/\sqrt{\hat{\lambda}_i}\right]}{\sum_{i=1}^{n}\left[(y_i - \bar{y})/\sqrt{\bar{y}}\right]}$$

  - Para una observación individual, la desviación del modelo permite obtener la suma de las desviaciones, las cuales se pueden utilizar como una medida de bondad del ajuste (siempre que se asuma que se incluye un término constante)

$$d_i = 2\left[y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right] = 2\left[y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - e_i\right]$$

$$where \;\; 0\ln(0) = 0$$

$$\Rightarrow G^2 = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - e_i\right] = 2\sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right)$$

- Cameron y Windmeijer, en su investigación de 1993, sugirieron el uso de una medida llamada R_d^2, la cual se puede expresar en términos de la función de la log-verosimilitud. Tanto el numerador como el denominador miden la mejora del ajuste del modelo sobre el modelo con un solo término constante, y el denominador mide la mejora máxima posible. Esta se define de la siguiente manera:

$$R_d^2 = 1 - \frac{\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - e_i\right]}{\sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{\bar{y}}\right)} = \frac{\ln L(\hat{\lambda}_i|y_i) - \ln L(\bar{y}|y_i)}{\ln L(\lambda_i|y_i) - \ln L(\bar{y}|y_i)} \in [0,1]$$

- Algunos programas utilizan una medida de bondad del ajuste parecida a R_d^2, llamada pseudo-R^2 o *likelihood ratio index*, la cual se define de la siguiente manera:

$$R_{pseudo}^2 = 1 - \frac{\ln L(\hat{\lambda}_i|y_i)}{\ln L(\bar{y}|y_i)}$$

- El modelo de Poisson se ha criticado porque hace una suposición implícita de que la varianza de y_i iguala a la media, el primer paso del análisis es contrastar esta suposición implícita en el contexto de un modelo simple

$$H_0 : Var(y_i) = E(y_i)$$

$$H_1 : Var(y_i) = E(y_i) + \alpha g[E(y_i)]$$

- Un contraste propuesto por Cameron y Trivedi se puede realizar a través de hacer una regresión para z_i, donde $\hat{\lambda}_i$ es el valor predicho de la regresión. Una vez hecho esto, un simple contraste t de si el vector de coeficientes es significativamente diferente de cero

$$z_i = \frac{(y_i - \hat{\lambda}_i)^2 - y_i}{\hat{\lambda}_i\sqrt{2}} \Rightarrow T = \frac{\boldsymbol{\beta}}{Std.Err(\boldsymbol{\beta})}$$

- Como la igualdad asumida entre la esperanza condicional y la función de varianza es una desventaja del modelo de Poisson, la alternativa más utilizada es la distribución negativa binomial, que nace de una formulación natural de heterogeneidad transversal

- Se generaliza el modelo de Poisson introduciendo un efecto individual no observable en la media condicional, en donde ε_i refleja el error de especificación o la heterogeneidad

$$\ln \mu_i = \boldsymbol{x_i'\beta} + \varepsilon_i = \ln \lambda_i + \ln u_i$$

- La distribución de y_i condicionada a $\boldsymbol{x_i}$ y u_i sigue siendo una distribución de Poisson con media condicional y varianza μ_i

$$f(y_i|\boldsymbol{x_i}, u_i) = \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!}$$

- La distribución incondicional $f(y_i|\boldsymbol{x_i})$ es el valor esperado de u_i de $f(y_i|\boldsymbol{x_i}, u_i)$, de modo que u_i define la distribución incondicional

$$f(y_i|\boldsymbol{x_i}) = \int_0^\infty \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!} g(u_i)\, du_i$$

- Por conveniencia matemática, se suele asumir una distribución gamma se asume para $u_i = \exp(\varepsilon_i)$. Como la media de la distribución no está identificada si el modelo contiene un término constante (porque las perturbaciones entran multiplicativamente), por lo que se asume que $E(u_i)$ y esta normalización causa que la función g sea de la siguiente manera:

$$g(u_i) = \frac{\theta^\theta}{\Gamma(\theta)} e^{-\theta u_i} u_i^{\theta-1}$$

- En consecuencia, la función de densidad para y_i con esta especificación es una forma de la distribución binomial negativa, la cual tiene media condicional λ_i y varianza $\lambda_i[1 + (1/\theta)\lambda_i]$

$$f(y_i|\boldsymbol{x_i}) = \int_0^\infty \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!} \frac{\theta^\theta}{\Gamma(\theta)} e^{-\theta u_i} u_i^{\theta-1}\, du_i =$$

$$= \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} r_i^{y_i}(1 - r_i)^\theta \quad where\ \ r_i = \frac{\lambda_i}{\lambda_i + \theta}$$

- Este modelo se denomina binomial negativo y puede ser estimado mediante máxima verosimilitud. Un contraste para la distribución de Poisson normalmente se lleva a cabo para la hipótesis $\alpha = 1/\theta = 0$ usando el contraste de Wald o de razón de verosimilitud

- Los efectos marginales, en este caso, son los mismos que los efectos marginales en el caso en donde la variable dependiente seguía una distribución de Poisson, debido a que la esperanza condicionada es la misma

$$\frac{\partial E(y_i|\boldsymbol{x}_i)}{\partial x_k} = \frac{\partial \lambda_i}{\partial x_k} = \beta_k \lambda_i = \beta_k e^{\boldsymbol{x}_i'\boldsymbol{\beta}}$$

- No se puede interpretar la magnitud del coeficiente (porque no es el coeficiente únicamente), pero si se puede interpretar el signo de los coeficientes

- El efecto marginal medio estimado, al ser la media de los efectos marginales estimados para cada individuo, también se puede expresar como el producto entre la media de la variable y y el coeficiente del regresor x_k para el que se estima el efecto

- Aunque la magnitud no sea interpretable en el efecto marginal, sí tiene una interpretación para el efecto sobre el valor esperado de λ_i igual que en el modelo de Poisson

- Con tal de poder modelar de manera más precisa el conteo de eventos con la presencia de muchos ceros (en relación a la cantidad esperada) se han utilizado modelos valla o *hurdle* y modelos inflados a cero o *zero-inflated models*

  - Algunos investigadores han analizado una extensión para el modelo valla en la que el resultado de cero puede producirse en dos regímenes

    - En un régimen, el resultado siempre es cero, mientras que, en el otro, el proceso de Poisson es el que puede producir ceros u otros valores (por lo que los ceros se pueden producir en ambos régimenes)

      $$P(y_i = 0|\boldsymbol{x_i}) = P(regime\ 1) + P(y_i = 0|\boldsymbol{x_i}, regime\ 2)P(regime\ 2)$$

      $$P(y_i = j|\boldsymbol{x_i}) = P(y_i = j|\boldsymbol{x_i}, regime\ 2)P(regime\ 2)\ \ for\ j = 1,2, \ldots$$

    - Siendo z un indicador binario del régimen (0 si se está en el primer régimen y 1 si está en el segundo régimen) y y^* el resultado del proceso de Poisson del segundo régimen, entonces la y observada es $z \times y^*$

  - Una extensión natural del modelo anterior es permitir que z se determine por un conjunto de variables que pueden ser iguales o no a las presentes en el proceso de Poisson. Esta extensión se llama modelo inflado a cero o *zero-inflated model* (ZIP)

    $$P(z_i = 0|\boldsymbol{w_i}) = F(\boldsymbol{w}_i'\boldsymbol{\gamma})\ \ (Regime\ 1)$$

    $$P(y_i = j|\boldsymbol{x_i}, z_i = 1) = \frac{\exp(-\lambda_i)\,\lambda_i^j}{j!}\ \ (Regime\ 2)$$

- El modelo inflado a cero puede verse como un tipo de modelo latente. Las dos clases de probabilidades son $F(\boldsymbol{w}_i, \boldsymbol{\gamma})$ y $1 - F(\boldsymbol{w}_i, \boldsymbol{\gamma})$ y los dos regímenes son $y = 0$ y el proceso de generación de datos Poisson. Se puede escoger una $F(\boldsymbol{w}_i, \boldsymbol{\gamma})$ *probit* o *logit*

- De manera alternativa, el modelo se puede expresar de la siguiente manera, dado que la formulación anterior no limita el dominio de *j*

$$P(y_i = 0|\boldsymbol{x_i}, \boldsymbol{w}_i, \boldsymbol{\gamma}) = F(\boldsymbol{w}_i, \boldsymbol{\gamma}) + [1 - F(\boldsymbol{w}_i, \boldsymbol{\gamma})]P(y_i = 0|\boldsymbol{x_i}, z_i = 1)$$

$$P(y_i = j|\boldsymbol{x_i}, \boldsymbol{w}_i, \boldsymbol{\gamma}) = [1 - F(\boldsymbol{w}_i, \boldsymbol{\gamma})]P(y_i = j|\boldsymbol{x_i}, z_i = 1) \ \ for \ j > 0$$

- Se puede hacer una extensión del modelo ZIP para incluir un proceso de generación de datos binomial negativo, llamado ZINB. En este modelo también se puede escoger una $F(\boldsymbol{w}_i, \boldsymbol{\gamma})$ *probit* o *logit*

- La media de la variable aleatoria en el caso de Poisson sería la siguiente:

$$E(y_i|\boldsymbol{x_i}, \boldsymbol{w}_i) = F_i \times 0 + (1 - F_i) \times E(y_i^*|\boldsymbol{x_i}, z_i = 1) = (1 - F_i)\lambda_i$$

- Aunque sería interesante contrastar si hay un régimen divisor o no, el modelo básico anterior y el modelo ZIP no son modelos anidados

  - No se puede contrastar si todos los coeficientes son nulos para comprobarlo, dado que eso no elimina la estructura divisoria, sino que modifica las probabilidades

  - Como se quiere un contraste que verifique si se sigue un proceso de Poisson o no (dado que otra distribución podría generar más ceros que una Poisson), esto hace que el contraste sea un procedimiento difícil y que las hipótesis no sean anidadas, haciendo que el poder del contraste sea una función de la hipótesis alternativa

- Vuong, en su investigación de 1989 propuso un estadístico de contraste para modelos no anidados que se adecúa a la situación en la que se puede especificar la distribución alternativa

  - Siendo $f_j(y_i|\boldsymbol{x_i})$ la probabilidad predicha de que la variable $Y = y_i$ bajo la suposición de que la distribución es $f_j(y_i|\boldsymbol{x_i})$ para $j = 1{,}2$ y definiendo $m_i = \ln[f_1(y_i|\boldsymbol{x_i})] - \ln[f_2(y_i|\boldsymbol{x_i})]$, entonces el estadístico para contrastar la hipótesis nula del modelo 1 respecto al 2 es el siguiente:

$$v = \frac{\frac{\sqrt{n}}{n}\sum_{i=1}^{n} m_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \overline{m})^2}} = \frac{\overline{m}\sqrt{n}}{s_m} \sim N(0,1)$$

- Este estadístico es bidireccional y permite contrastar la hipótesis de que $E(m_i) = 0$. Si $|v| < 2$ entonces el contraste no favorece a ningún modelo, pero para valores absolutos grandes, se apoya un modelo u otro (al 1 si es positivo o al 2 si es negativo)

- Rechazar la hipótesis nula quiere decir rechazar que el proceso generador de datos es una Poisson común, de modo que el modelo ZIP se adecúa mejor a los datos. Se puede hacer del mismo modo con una distribución binomial negativa

- No obstante, existe un debate sobre el uso de este contraste para contrastar los modelos ZIP contra los de conteo comunes, debido a la definición y prerrequisitos que determina Vuong para su uso

- En algunos contextos, el resultado cero del proceso generador de datos es cualitativamente diferente de los valores positivos, de modo que se utilizan modelos valla o *hurdle models*

  - Este es el caso cuando el resultado proviene de una decisión separada sobre si participar o no en la actividad y, al decidir si participar o no, el individuo hace una decisión sobre cuánto o sobre la intensidad

$$P(y_i = 0|\boldsymbol{w_i}) = F(\boldsymbol{w}_i, \boldsymbol{\gamma})$$

$$P(y_i = j|\boldsymbol{x_i}, \boldsymbol{w_i}, y_i > 0) = [1 - F(\boldsymbol{w}_i, \boldsymbol{\gamma})]\frac{\exp(-\lambda_i)\,\lambda_i^j}{j!\,[1 - \exp(-\lambda_i)]}$$

  - Como se puede ver en esta formulación, los ceros se determinan en la primera ecuación, pero no en la segunda, de modo que se cambia la probabilidad de obtener un resultado nulo y se escalan las probabilidades que sobran para que sumen a uno

  - Igual que antes, se puede escoger una forma funcional para $F(\boldsymbol{w}_i, \boldsymbol{\gamma})$ que sea binaria como el *logit* o el *probit* y se puede extender el modelo a utilizar un proceso binomial negativo en la segunda ecuación

  - La media condicional del modelo valla con un proceso de Poisson permite obtener fácilmente los efectos parciales

$$E[y_i|\boldsymbol{x}_i, \boldsymbol{w}_i] = \frac{[1 - F(\boldsymbol{w}_i'\boldsymbol{\gamma})]\lambda_i}{1 - \exp(-\lambda_i)} \quad where \quad \lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$$

$$\frac{\partial E[y_i|\boldsymbol{x}_i, \boldsymbol{w}_i]}{\partial \boldsymbol{x}_i} = [1 - F(\boldsymbol{w}_i'\boldsymbol{\gamma})]\boldsymbol{\delta_i}$$

$$\frac{\partial E[y_i|\boldsymbol{x}_i, \boldsymbol{w}_i]}{\partial \boldsymbol{w}_i} = \left[-\frac{f(\boldsymbol{w}_i'\boldsymbol{\gamma})}{1 - \exp(-\lambda_i)}\right]\boldsymbol{\gamma}$$

        - Para aquellas variables que aparecen tanto en $\boldsymbol{x}_i$ como en $\boldsymbol{w}_i$, los efectos se suman entre sí para obtener el efecto total de esa variable. Para variables binarias, en cambio, las fórmulas dan un resultado aproximado, dado que se tendrían que tomar las diferencias de las medias condicionales cuando la variable pasa de cero a uno

        - Puede ser de interés contrastar los efectos valla, pero igual que con el modelo ZIP, el modelo valla tiene hipótesis no anidadas (produciendo los mismos resultados que antes en la probabilidad), de modo que se suele utilizar el contraste de Vuong visto anteriormente en el contexto del modelo ZIP

    - El modelo valla guarda similitudes con el ZIP, pero las implicaciones del comportamiento de los individuos en los modelos son diferentes y las modificaciones alteran la formulación de Poisson o de la binomial negativa

        - Ambos modelos tienen sobredispersión, por lo que no hay una igualdad entre la media y la varianza como en el proceso de Poisson. No obstante, esta sobredispersión no nace de la heterogeneidad, sino del proceso generador de ceros

        - Esto provoca un problema de identificación, dado que es difícil saber si el exceso de ceros en los datos se produce por la heterogeneidad o por el mecanismo divisor

Los modelos de truncación y censura

- Muchas veces se tiene interés en analizar variables que no pueden ser observadas directamente porque estas muestran información incompleta porque están truncadas o censuradas

    - La truncación y la censura dificultan el análisis de datos que, de otro modo, se podrían analizar a través de modelos econométricos más simples como las regresiones

        - Se puede considerar que la censura y la truncación son situaciones en las que no se tiene información completa sobre una variable latente Y^* de interés (solo de la variable Y)

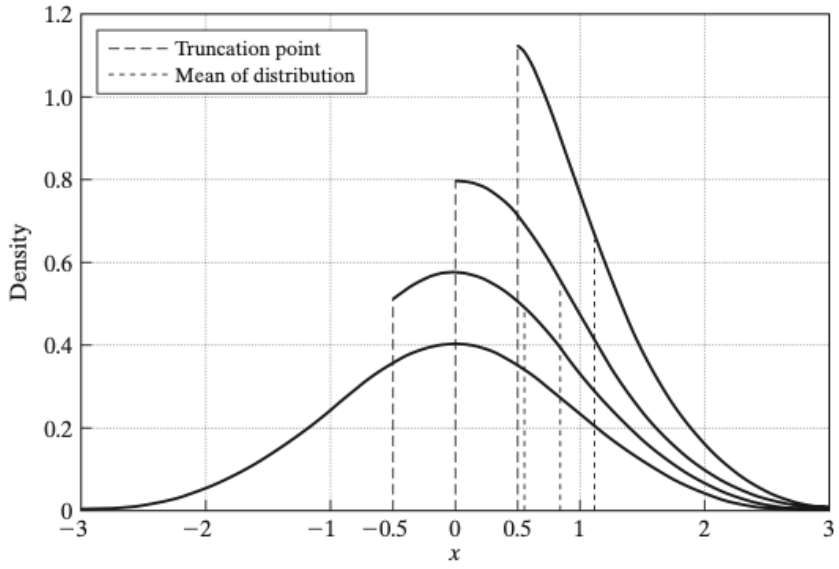
    - La truncación nace como consecuencia de descartar datos, tales como valores nulos, negativos o inusuales. La censura, en cambio, es una característica del diseño de muestreo en el que se representan ciertos valores con otros valores (de modo que no se saben los valores originales)

        - La truncación es esencialmente una característica de la distribución de la que se coge la muestra, mientras que la censura es un defecto en los datos muestrales, dado que, si no se censurara, los datos serían una muestra representativa de la población

        - La truncación y la censura producen unos sesgos sistemáticos al hacer inferencias sobre la población entera, llamados sesgo de truncación o *truncation bias* y sesgo de truncación o *truncation bias*

    - La truncación y la censura producen efectos similares en la distribución de valores de las variables aleatorias y en características como la media de esta distribución

- Una distribución truncada es la parte de una distribución no truncada que está por encima o por debajo de un valor especifico, por lo que se coge como datos un subconjunto de la distribución entera



    - Si una variable aleatoria x continua tiene una función de densidad de probabilidad $f(x)$ y a es una constante, entonces la función de

densidad de la variable x truncada será una función de densidad condicional para valores mayores a a será la siguiente:

$$f(x|x > a) = \frac{f(x)}{P(x > a)}$$

- La demostración proviene de la definición de la probabilidad condicional y se basa en escalar la densidad para que al integrar la función se obtenga 1 en el rango de a (para cumplir con las propiedades de la probabilidad)

- Una función muy similar aplica para el caso en que la truncación no es por abajo ($x > a$) si no por arriba ($x < a$)

- Normalmente se utiliza la distribución normal truncada en aplicaciones recientes. Si la variable x sigue una distribución normal con media μ y desviación estándar σ, entonces:

$$f(x|x > a) = \frac{f(x)}{1 - \Phi(\alpha)} = \frac{\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)}{1 - \Phi(\alpha)} \quad where \quad \alpha = \frac{a-\mu}{\sigma}$$

- La probabilidad de que $x > a$ cuando x se distribuye normalmente es $1 - \Phi(\alpha)$

$$P(x > a) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right) = 1 - \Phi(\alpha)$$

- Se estandarizan la variable x y el valor a porque se utiliza la función de distribución y de densidad de una variable normal estándar. Es posible utilizar una distribución normal estándar truncada, por lo que no se estandarizarían los valores

$$P(x > a) = 1 - \Phi(a)$$

$$f(x|x > a) = \frac{f(x)}{1 - \Phi(a)} = \frac{\frac{1}{\sigma}\phi(x)}{1 - \Phi(a)}$$

- Las fórmulas anteriores se alteran ligeramente para variables truncadas por arriba, dado que la probabilidad cambia

$$P(x < a) = \Phi\left(\frac{a-\mu}{\sigma}\right) = \Phi(\alpha)$$

- Normalmente, uno se interesa por la media y la varianza de una variable aleatoria truncada, las cuales se obtendrían con las siguientes fórmulas generales:

$$E(x|x > a) = \int_a^{\infty} xf(x|x > a)\, dx$$

$$Var(x|x > a) = \int_a^{\infty} [x - E(x|x > a)]^2 f(x|x > a)\, dx$$

- Los momentos de una distribución normal se pueden obtener del mismo modo, resultando en las siguientes fórmulas:

$$E(x|x > a) = \mu + \sigma\lambda(\alpha)$$

$$Var(x|x > a) = \sigma^2[1 - \delta(\alpha)]$$

$$where\ \ \lambda(\alpha) = \begin{cases} \dfrac{\phi(\alpha)}{1 - \Phi(\alpha)} & if\ truncation\ is\ x > a \\ -\dfrac{\phi(\alpha)}{\Phi(\alpha)} & if\ truncation\ is\ x < a \end{cases}$$

$$and\ \ \ \delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha]$$

- La variable $\lambda(\alpha)$ se denomina *ratio* inversa de Mills, y se define como el cociente entre la función de densidad normal y la función de distribución de probabilidad normal. Esta función es positiva en todo su dominio, y por definición, se cumple que:

$$-\frac{1}{2\sqrt{2\pi}} < \lambda(\alpha) < \frac{1}{2\sqrt{2\pi}} \ \Rightarrow\ 0 < \delta(\alpha) < 1$$

- La truncación tiene dos efectos importantes en los momentos estadísticos de la distribución de la variable:

  - Si la truncación es por abajo, la media será más alta que la media de la distribución original, mientras que, si la truncación es por arriba, la media será menor

  - Como $0 < \delta(\alpha) < 1$, la varianza de la distribución será menor que la varianza original, dado que incluye menos valores extremos

- A partir del concepto de truncación y de las propiedades de la distribución de una variable aleatoria truncada, es posible crear un modelo de regresión para lidiar con esta, llamado modelo de regresión truncada

  - Se asume que hay una variable Y_i^* latente, la cual está determinada por unos regresores $\boldsymbol{X_i}$ y un término de error ε_i

$$Y_i^* = \boldsymbol{x_i'\beta} + \varepsilon_i$$

- Se asume que el término de error está normalmente distribuido con media nula y varianza σ^2, de modo que la variable Y_i^* también sigue una distribución normal

$$\varepsilon_i \sim N(0, \sigma^2) \;\Rightarrow\; Y_i^* \sim N(\boldsymbol{x_i'\beta}, \sigma^2)$$

- El foco principal está en la distribución de la variable truncada Y_i, la cual dependerá de la variable latente Y_i^* y de su esperanza, que es una función no lineal de a, σ, $\boldsymbol{x}$ y $\boldsymbol{\beta}$

$$E(Y_i^*|\boldsymbol{X_i}; Y_i^* > a) = \boldsymbol{x_i'\beta} + \sigma\lambda(\alpha_i) = \boldsymbol{x_i'\beta} + \sigma\frac{\phi(\alpha_i)}{1-\Phi(\alpha_i)}$$

$$where\ \ \alpha_i = \frac{a - \boldsymbol{x_i'\beta}}{\sigma}$$

  - En este caso, la media de Y_i^* es $\boldsymbol{X_i'\beta}$, por lo que α_i dependerá del individuo y la esperanza también

  - La esperanza se puede expresar de manera similar cuando la variable está truncada por arriba

$$E(Y_i^*|\boldsymbol{x_i}; Y_i^* > a) = \boldsymbol{x_i'\beta} + \sigma\lambda(\alpha_i) = \boldsymbol{x_i'\beta} - \sigma\frac{\phi(\alpha_i)}{\Phi(\alpha_i)}$$

- A través de la esperanza condicional de Y_i^*, es posible obtener el efecto marginal de los regresores para una variable truncada y el efecto en la varianza

$$\frac{\partial E(Y_i^*|\boldsymbol{x_i}; Y_i^* > a)}{\partial X_{ki}} = \beta_k + \sigma\frac{\partial\lambda(\alpha_i)}{\partial\alpha_i}\frac{\partial\alpha_i}{\partial X_{ki}} =$$

$$= \beta_k + \sigma\left[\frac{\phi(\alpha_i)^2}{[1-\Phi(\alpha_i)]^2} - \frac{\alpha_i\phi(\alpha_i)}{1-\Phi(\alpha_i)}\right]\left(\frac{\beta_k}{\sigma}\right) =$$

$$= \beta_k[1 + \lambda(\alpha_i)^2 - \alpha_i\lambda(\alpha_i)] = \beta_k[1 - \delta(\alpha_i)]$$

$$Var(Y_i^*|\boldsymbol{x_i}, Y_i^* > a) = \sigma[1 - \delta(\alpha_i)]$$

  - Para $Y_i^* < a$, se obtiene una expresión equivalente que resulta en el mismo efecto marginal

  - El efecto en Y_i de un incremento unitario de una variable X_{ki} se debe interpretar como el efecto que tiene sobre aquellos individuos que previamente tenían un valor dentro de la

distribución truncada, dado que solo se tienen esas observaciones en la muestra y no se puede generalizar el efecto para aquellos individuos que no estén dentro (no se tienen en cuenta las otras observaciones)

- Debido a que $0 < \delta(\alpha_i) < 1$, el efecto marginal de aumentar una unidad en el regresor X_{ki} es menor a su coeficiente en valo absoluto, y la varianza es menor a la que tendría la variable latente, por lo que hay una atenuación debido a la truncación

$$|ME_{trun}| \leq |\beta_k|$$

- Cuando un regresor está en el modelo con una forma polinómica o con una interacción, el efecto marginal no siempre tiene el mismo signo que el primer coeficiente del regresor, debido a que tiene dos o más y depende de los signos de estos coeficientes

- Si el estudio econométrico se centra en la población, entonces los coeficientes $\boldsymbol{\beta}$ son los elementos de interés, dado que el efecto marginal obtenido es el de la subpoblación, pero el efecto marginal de las variables latentes serían los coeficientes

- Para la subpoblación de la que se obtienen los datos de la variable latente, se puede utilizar un modelo de regresión de la siguiente forma:

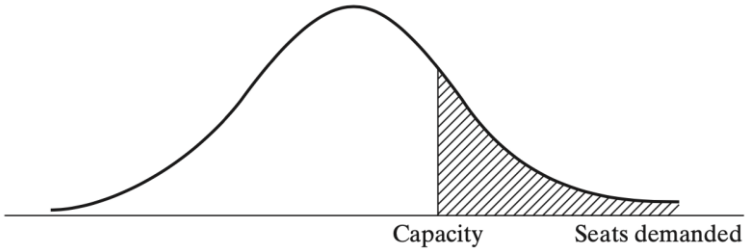
$$Y_i = E(Y_i^* | \boldsymbol{x_i}, Y_i^* > a) = \boldsymbol{x_i'\beta} + \sigma\lambda(\alpha_i) + u_i$$

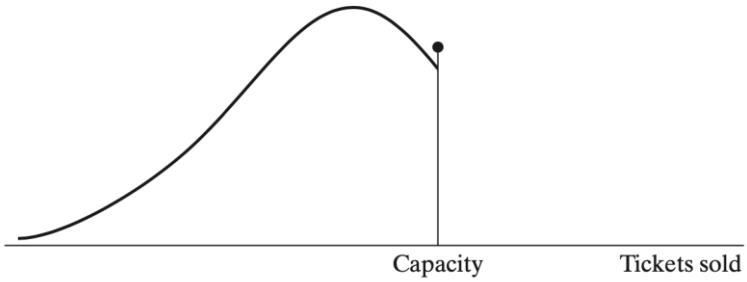
$$where\ u_i = Y_i^* - E(Y_i^* | \boldsymbol{x_i}, Y_i^* > a)$$

- Por definición, el término de error u_i se distribuirá de manera normal con media nula, pero tiene una varianza equivalente a $Var(Y_i^* | \boldsymbol{x_i}, Y_i^* > a)$ y es heteroscedástico (dado que la varianza de la variable truncada depende de $\boldsymbol{x_i}$)

- Si se estima el modelo por mínimos cuadrados ordinarios, no obstante, los estimadores estarán sesgados porque se ignora el término no lineal $\lambda(\alpha_i)$ y habría un sesgo por omisión de variables

- La teoría relevante para la distribución de una variable censurada es similar a la de una truncada, por lo que los aspectos relevantes son similares





- Cuando los datos están censurados, la distribución que aplica a la muestra de estos datos es una combinación de distribuciones discretas y continuas. Para analizar esta distribución, se define una variable Y_i transformada a partir de Y_i^*

$$Y_i = \begin{cases} a & if \;\; Y_i^* \geq a \\ Y_i^* & if \;\; Y_i^* < a \end{cases}$$

  - La distribución que aplica cuando $Y_i^* \leq a$ es la función de distribución de probabilidad $P(Y_i = a) = P(Y_i^* \geq a)$, mientras que cuando $Y^* < a$, la función de densidad será $f(Y) = f(Y_i^*)$. De este modo, la probabilidad total es igual a 1 y se le asigna la probabilidad de la parte censurada al punto de censura a

  - La censura mostrada en el ejemplo es una censura a la derecha, pero también se puede censurar a la izquierda y censurar tanto a la izquierda como a la derecha

$$Y_i = \begin{cases} Y_i^* & if \;\; Y^* < a \\ a & if \;\; Y_i^* \geq a \end{cases} \qquad Y_i = \begin{cases} a & if \;\; Y_i^* < a \\ Y_i^* & if \;\; a \leq Y_i^* \leq b \\ b & if \;\; Y_i^* > b \end{cases}$$

  - Para una variable truncada, la única parte relevante para los cálculos de su distribución es la parte donde $Y_i^* > a$ o $Y_i^* < a$, de modo que se escala la función de densidad con la probabilidad para que la integral sea igual a 1

- Se suele asumir que la variable aleatoria censurada sigue una distribución normal, dado que es lo que más se usa en las investigaciones. Por lo tanto, se pueden analizar los momentos estadísticos de la distribución censurada bajo esta suposición

$$If\ \ Y_i^* \sim N(\mu, \sigma^2)\ \ and\ \ \alpha = \frac{a-\mu}{\sigma}:$$

$$E(Y_i) = \Phi(\alpha)a + [1-\Phi(\alpha)][\mu + \sigma\lambda(\alpha)]$$

$$Var(Y_i) = \sigma^2[1-\Phi(\alpha)]\big[[1-\delta(\alpha)] + [\alpha - \lambda(\alpha)]^2\Phi(\alpha)\big]$$

- Las funciones de la *ratio* inversa de Mills y de la distribución normal se definen de la manera ya mencionada anteriormente con las variables truncadas. Para cada tipo de censura, se redefine la forma funcional de la probabilidad (dado que varía)

- La demostración de la esperanza bajo la suposición de distribución normal es la siguiente:

$$E(Y_i) = E(Y_i|Y_i < a)P(Y_i < a) =$$

$$= P(Y_i = a)E(Y_i|Y_i = a) + P(Y_i < a)E(Y_i|Y_i < a) =$$

$$= P(Y_i^* \geq a)a + P(Y_i^* < a)E(Y_i^*|Y_i^* < a) =$$

$$= P(Y_i^* \geq a)a + P(Y_i^* < a)[\mu + \sigma\lambda(\alpha)]$$

- Para el caso especial en que $a = 0$, los momentos se pueden expresar de la siguiente manera:

$$E(Y_i) = \left[1 - \Phi\left(-\frac{\mu}{\sigma}\right)\right]\left[\mu + \sigma\lambda\left(-\frac{\mu}{\sigma}\right)\right]$$

$$Var(Y_i) = \sigma^2\left[1 - \Phi\left(-\frac{\mu}{\sigma}\right)\right]\left[\left[1 - \delta\left(-\frac{\mu}{\sigma}\right)\right] + \left[-\frac{\mu}{\sigma} - \lambda\left(-\frac{\mu}{\sigma}\right)\right]^2 \Phi\left(-\frac{\mu}{\sigma}\right)\right]$$

  - Una variable puede estar censurada y truncada a la vez, de modo que hay valores de Y^* para los cuales se censura con un valor concreto, pero hay valores que quedan fuera de la muestra:

$$Y = \begin{cases} Y^* & if\ \ b \leq Y^* \leq c \\ b & if\ \ a < Y^* < b \end{cases}$$

    - En este caso en concreto, la variable está truncada de b a c (valores por encima de c o por debajo de b no entran en la muestra), pero está censurada a la izquierda de b (los valores entre a y b tienen un valor $Y = b$)

- A partir del concepto de censura y de las propiedades de la distribución de una variable aleatoria censurada, es posible crear un modelo de regresión para lidiar con esta, llamado modelo Tobit o de regresión censurada

- Se asume que hay una variable Y_i^* latente, la cual está determinada por unos regresores $\boldsymbol{X_i}$ y un término de error ε_i, y que hay una variable observada Y_i que está censurada

$$Y_i^* = \boldsymbol{x_i'\beta} + \varepsilon_i \qquad Y_i = \begin{cases} a & if \;\; Y_i^* \geq a \\ Y_i^* & if \;\; Y_i^* < a \end{cases}$$

  - La variable observada se puede censurar de diferentes maneras, de modo que se puede definir de maneras alternativas

  - Se asume que el término de error está normalmente distribuido con media nula y varianza σ^2, de modo que la variable Y_i^* también sigue una distribución normal

$$\varepsilon_i \sim N(0, \sigma^2) \;\; \Rightarrow \;\; Y_i^* \sim N(\boldsymbol{x_i'\beta}, \sigma^2)$$

- El foco principal está en la distribución de la variable censurada Y_i, la cual dependerá de la variable latente Y_i^* y de su esperanza, que es una función no lineal de a, σ, $\boldsymbol{x}$ y $\boldsymbol{\beta}$

$$E(Y_i|\boldsymbol{X_i}) = \Phi(\alpha_i)a + [1 - \Phi(\alpha_i)][\boldsymbol{x_i'\beta} + \sigma\lambda(\alpha_i)]$$

$$where \;\; \alpha_i = \frac{a - \boldsymbol{x_i'\beta}}{\sigma}$$

  - En este caso, la media de Y_i^* es $\boldsymbol{X_i'\beta}$, por lo que α_i dependerá del individuo y la esperanza también

  - La definición de la *ratio* inversa de Mills y las probabilidades de cada valor dependerán de la definición de la variable censurada, de modo que el resultado obtenido es generalizable a más casos

  - En el caso especial en el que $a = 0$, se puede obtener la siguiente expresión, la cual también es generalizable:

$$E(Y_i|\boldsymbol{X_i}) = \left[1 - \Phi\left(\frac{-\boldsymbol{x_i'\beta}}{\sigma}\right)\right]\left[\boldsymbol{x_i'\beta} + \sigma\lambda\left(\frac{-\boldsymbol{x_i'\beta}}{\sigma}\right)\right]$$

$$= \Phi\left(\frac{\boldsymbol{x_i'\beta}}{\sigma}\right)\left[\boldsymbol{X_i'\beta} + \sigma\lambda\left(\frac{-\boldsymbol{x_i'\beta}}{\sigma}\right)\right]$$

- A través de la esperanza condicional de Y_i, es posible obtener el efecto marginal de los regresores para una variable censurada

$$\frac{\partial E(Y_i|\boldsymbol{x_i})}{\partial X_{ki}} = \beta_k P(a < Y_i^* < b|\boldsymbol{x_i}) = \beta_k \left[F\left(\frac{b - \boldsymbol{x_i'\beta}}{\sigma}\right) - F\left(\frac{a - \boldsymbol{x_i'\beta}}{\sigma}\right)\right]$$

$$where\ F\ is\ the\ cond.prob.distrib.function\ of\ Y_i$$

- Esta expresión es generalizable para casos en los que no se censura entre dos valores y para casos en los que la distribución no sea normal (los errores no se distribuyen normalmente)

- El efecto en Y_i (siendo censurada) de un incremento unitario de una variable X_{ki} se debe interpretar como el efecto que tiene sobre aquellos individuos que tienen valores dentro de la distribución censurada y aquellos que tienen valores fuera de ella (el efecto se interpreta para la variable latente)

- Debido a que la probabilidad está entre $[0,1]$, el efecto marginal de aumentar una unidad en el regresor X_{ki} es siempre menor o igual a su coeficiente en valor absoluto

$$|ME_{cens}| \leq |\beta_k|$$

- Cuando un regresor está en el modelo con una forma polinómica o con una interacción, el efecto marginal no siempre tiene el mismo signo que el primer coeficiente del regresor, debido a que tiene dos o más y depende de los signos de estos coeficientes

- McDonald y Moffitt sugieren una descomposición del efecto marginal del modelo Tobit que permite analizar el efecto de un aumento unitario en un regresor

$$\frac{\partial E(Y_i|\boldsymbol{x_i})}{\partial X_{ki}} = \beta_k\big[\Phi(\alpha_i)[1-\delta(\alpha_i)] + \phi(\alpha_i)[\alpha_i + \lambda(\alpha_i)]\big] =$$

$$= P(Y_i < a|\boldsymbol{x_i})\frac{\partial E(Y_i|\boldsymbol{x_i}; Y_i < a)}{\partial X_{ki}} + E(Y_i|\boldsymbol{x_i}; Y_i < a)\frac{\partial P(Y_i < a|\boldsymbol{x_i})}{\partial X_{ki}}$$

  - La primera parte de la expresión es el efecto en el valor esperado condicional en la parte no censurada de la distribución, la cual se llama margen intensivo. La segunda parte es el efecto en la probabilidad de que la observación pertenezca a esa parte de la distribución, llamada margen extensivo

$$ME_{cens} = ME_{int} + ME_{ext}$$

$$ME_{int} = P(Y_i < a|\boldsymbol{x_i})\frac{\partial E(Y_i|\boldsymbol{x_i}; Y_i < a)}{\partial X_{ki}}$$

$$ME_{ext} = E(Y_i|\boldsymbol{x_i}; Y_i < a)\frac{\partial P(Y_i < a|\boldsymbol{x_i})}{\partial X_{ki}}$$

- Como el efecto marginal de un regresor en la variable censurada debe ser menor o igual en valor absoluto al coeficiente, los márgenes tienen que se ambos menores al valor absoluto del coeficiente

$$|ME_{cens}| \leq |\beta_k| \Rightarrow |ME_{int}| < |\beta_k| \ and \ |ME_{ext}| < |\beta_k|$$

- A través de analizar el margen intensivo y el margen extensivo, se puede ver como un incremento en el regresor tiene un efecto en la esperanza condicional y en la probabilidad de pertenecer a la distribución censurada con el mismo signo (al ser múltiplos positivos de β_k)

$$\frac{\partial P(Y_i < a|\boldsymbol{x_i})}{\partial X_{ki}} = \left(\frac{\beta_k}{\sigma}\right)\phi(\alpha_i)$$

$$\frac{\partial E(Y_i|\boldsymbol{x_i}; Y_i < a)}{\partial X_{ki}} = \left(\frac{\beta_k}{\sigma}\right)[1 - \delta(\alpha_i)]$$

- La proporción que representa el margen extensivo y el margen intensivo en el efecto marginal censurado debe ser el mismo para todos los individuos, aunque el efecto marginal censurado difiera para cada uno

- Para la población de la que se obtienen los datos de la variable latente, se puede utilizar un modelo de regresión con la esperanza, llamado modelo Tobit:

$$Y_i = E(Y_i|\boldsymbol{x_i}) = \Phi(\alpha_i)a + [1 - \Phi(\alpha_i)][\boldsymbol{x_i'\beta} + \sigma\lambda(\alpha_i)] + u_i$$

$$where \ u_i = Y_i - E(Y_i|\boldsymbol{x_i})$$

- Por definición, el término de error u_i se distribuirá de manera normal con media nula, pero tiene una varianza equivalente a $Var(Y_i|\boldsymbol{x_i})$ y es heteroscedástico (dado que la varianza de la variable censurada depende de $\boldsymbol{x_i}$)

- Si se estima el modelo por mínimos cuadrados ordinarios, no obstante, los estimadores estarán sesgados porque se ignora el término no lineal $\lambda(\alpha_i)$ y habría un sesgo por omisión de variables

- El modelo de Tobit tiene algunas limitaciones debido a la suposiciones que se hacen y a su especificación

- Se suelen hacer las suposiciones de normalidad y homoscedasticidad, pero estas pueden no ser apropiadas dada la población y la muestra

- El modelo Tobit propone un modelo para la variable observada Y_i en base a la variable latente Y_i^*. No obstante, Y_i depende de una variable endógena implícita, la cual se puede entender como la pertenencia a la distribución censurada o no (una variable binaria D_i), y como no está modelada, se asume que el efecto de las $\boldsymbol{x_i}$ es el mismo para ambas, lo cual es poco realista (no se participa porque no se quiere)

- Para estimar los coeficientes mediante el método de máxima verosimilitud se tiene que maximizar la función de verosimilitud en relación a los coeficientes de los regresores y al parámetro σ

  - La función de verosimilitud para una variable censurada a la derecha será la siguiente:

$$L = \prod_{i=1}^{N_1} P(Y_i = a) \prod_{i=N_1+1}^{N} f(Y_i) = \prod_{i=1}^{N_1} [1 - \Phi(\alpha_i)] \prod_{i=N_1+1}^{N} \frac{\phi\left[\frac{Y_i}{\sigma} - \alpha_i\right]}{\sigma}$$

$$1, \dots, N_1 = censored\ individuals$$

$$N_1 + 1, \dots, N = uncensored\ individuals$$

    - Como en realidad hay dos grupos de individuos, se pueden dividir entre aquellos con un valor positivo de Y_i^* y aquellos con un valor de a

    - En el grupo de aquellos que no consumen solo se puede observar un valor discreto a, de modo que se puede utilizar la probabilidad puntual (condicionada a los valores de los regresores de cada individuo) para expresar la probabilidad de que $Y = a$ (de que $Y_i^* \geq a$)

$$P(Y_i^* < a) + P(Y_i^* \geq a) = \Phi(\alpha_i) + 1 - \Phi(\alpha_i) = 1$$

$$P(Y_i = a) = P(Y_i^* \geq a) = 1 - \Phi(\alpha_i) = \Phi(-\alpha_i)$$

    - En el grupo de aquellos que consumen se puede observar infinitos valores positivos continuos, de modo que se tiene que utilizar la función de densidad de probabilidad condicionada a los valores de los regresores de cada individuo

- La función de verosimilitud para una variable truncada a la derecha, en cambio, será la siguiente:

$$L = \prod_{i=1}^{N} \frac{f(Y_i)}{P(Y < a)} = \prod_{i=1}^{N} \frac{\frac{1}{\sigma}\,\phi\left[\frac{Y_i^*}{\sigma} - \alpha_i\right]}{\Phi(\alpha_i)}$$

  - En este caso, la muestra solo incluye observaciones para las cuales $Y < a$, de modo que no hay dos grupos como en el caso anterior y todos los individuos pertenecen al mismo (no es una muestra representativa de la población porque solo representa la parte de la población con valores positivos)

  - Como solo se tienen valores superiores a a, la probabilidad de que se de un valor positivo cualquiera de Y_i (condicionada a los valores de los regresores de cada individuo) está condicionada a la probabilidad (condicionada) de que $Y_i^* < a$ (por la representatividad de la muestra), de modo que se divide entre la probabilidad puntual de que $Y_i^* < a$

  - Dado que la probabilidad puntual condicionada de que $Y_i^* < a$ sumada a la probabilidad puntual condicionada de que $Y_i = a$ es toda la población, entonces:

$$P(Y_i^* < a) + P(Y_i^* \geq a) = \Phi(\alpha_i) + 1 - \Phi(\alpha_i) = 1$$

$$P(Y_i < a) = P(Y_i^* < a) = \Phi(\alpha_i) = 1 - \Phi(-\alpha_i)$$

- Se maximiza la función de verosimilitud en base a los coeficientes y a la σ con tal de poder obtener los coeficientes que maximizan la función:

$$\max_{\boldsymbol{\beta},\sigma} L = \prod_{i=1}^{N_1} [1 - \Phi(\alpha_i)] \prod_{i=N_1+1}^{N} \frac{\phi\left[\frac{Y_i}{\sigma} - \alpha_i\right]}{\sigma}$$

$$\max_{\boldsymbol{\beta},\sigma} L = \prod_{i=1}^{N} \frac{\frac{1}{\sigma}\,\phi\left[\frac{Y_i}{\sigma} - \alpha_i\right]}{\Phi(\alpha_i)}$$

  - En este caso, el sistema de ecuaciones con las condiciones de primer orden para todos los coeficientes también tiene una ecuación de primer orden para la desviación típica σ, de modo que también se puede estimar

$$\boldsymbol{\beta}, \sigma \;\Rightarrow\; \widehat{\boldsymbol{\beta}}, \hat{\sigma}$$

- Los estimadores obtenidos por máxima verosimilitud solo serán eficientes si el error se distribuye de manera normal y los errores son homoscedásticos (además de si la especificación del modelo es buena)

- La función de verosimilitud para una variable censurada y truncada a la vez tendrá que dividir entre la probabilidad que representa la muestra, pero teniendo en cuenta la probabilidad de ambos valores de Y:

$$L = \prod_{i=1}^{N_1} \frac{P(a < Y^* < b)}{P(a < Y^* < c)} \prod_{i=N_1+1}^{N} \frac{f(Y^*)}{P(a < Y^* < c)}$$

- Para estimar los coeficientes de ambos modelos sin sesgos, se puede usar un método de dos etapas en donde se hace una primera estimación con un modelo *probit* (primera etapa) y una segunda estimación (segunda etapa) mediante mínimos cuadrados ordinarios

  - Los individuos se pueden dividir en dos grupos cuando se trunca o se censura una variable: los que tienen valores $Y_i^* > a$ y aquellos que no. Por lo tanto, se puede crear una variable binaria para usarse como dependiente en un modelo *probit*

$$D_i = \begin{cases} 1 \ \ if \ Y_i^* > a \\ 0 \ \ if \ Y_i^* \leq a \end{cases} \Rightarrow \ d_i^* = \boldsymbol{x_i'\beta} + u_i \quad where \quad u_i \sim N(0,1)$$

$$\Rightarrow \ P(D_i = 1|\boldsymbol{x_i}) = \Phi(\alpha_i)$$

    - Estos coeficientes son diferentes de las $\boldsymbol{\beta}$ y se estiman por máxima verosimilitud

    - Se escoge una definición de D_i u otra dependiendo de si la truncación o la censura es por arriba, por abajo, por la derecha, por la izquierda o por ambas

  - Como la probabilidad de que $D_i = 1$ es igual a la probabilidad de que $d_i^* > a$, se puede ver que existe una relación entre los coeficientes del modelo de truncación o censura y los del modelo *probit* si el modelo es correcto

$$P(d_i^* > a) = 1 - \Phi(a - \boldsymbol{x_i'\beta}) \ \Rightarrow \ \Phi(\boldsymbol{x_i'\gamma}) = \Phi(\boldsymbol{x_i'\beta} - a)$$

    - Se puede ver como hay una relación de proporcionalidad entre los coeficientes $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ de los regresores no constantes: los coeficientes del modelo no son más que los coeficientes del *probit* multiplicados por σ

$$\boldsymbol{\gamma} = \frac{\boldsymbol{\beta}}{\sigma} \Rightarrow \boldsymbol{\beta} = \boldsymbol{\gamma}\sigma$$

- Parar el coeficiente constante, la relación será la misma, pero se le añadirá o se le sustraerá a dependiendo de la especificación de la variable dependiente observada

$$\gamma_0 = \frac{\beta_0 \pm a}{\sigma} \Rightarrow \beta = \gamma\sigma \pm a$$

- La segunda etapa para una variable truncada consiste en hacer una regresión lineal de la variable dependiente Y_i sobre los regresores, de modo que se estiman los coeficientes de dos etapas por mínimos cuadrados

$$\widehat{Y}_\iota = E(Y_i|\boldsymbol{x_i}; D_i = 1) = \boldsymbol{x_i'}\widehat{\boldsymbol{\beta}} + \hat{\sigma}\lambda(\widehat{\alpha_\iota})$$

$$By\ MCO\colon\ \widehat{\boldsymbol{\beta}}^{2stage}, \hat{\sigma}^{2stage}$$

- El valor de $\lambda(\widehat{\alpha_\iota})$ (que funciona como regresor) se obtiene con los coeficientes de la primera parte

- La segunda etapa para una muestra censurada consiste en hacer una regresión lineal de la variable dependiente Y_i sobre los regresores multiplicados por $\Phi(\widehat{\alpha_\iota})$, de modo que se estiman los coeficientes de dos etapas por mínimos cuadrados

$$\widehat{Y}_\iota = E(Y_i|\boldsymbol{x_i};) = \Phi(\widehat{\alpha_\iota})a + [1 - \Phi(\widehat{\alpha_\iota})]\left[\boldsymbol{x_i'}\widehat{\boldsymbol{\beta}} + \hat{\sigma}\lambda(\widehat{\alpha_\iota})\right]$$

$$By\ MCO\colon\ \widehat{\boldsymbol{\beta}}^{2stage}, \hat{\sigma}^{2stage}$$

- El valor de $\Phi(\widehat{\alpha_\iota})$ y $\phi(\widehat{\alpha_\iota})$ se obtienen con los coeficientes de la primera parte, y $\phi(\widehat{\alpha_\iota})$ funciona como un regresor más

- Para estimar por mínimos cuadrados, sin embargo, hace falta multiplicar los valores de $\Phi(\widehat{\alpha_\iota})$ por los de $\boldsymbol{x}$ (creando así nuevos regresores) para obtener los coeficientes de $\boldsymbol{x}$. El coeficiente constante, por tanto, ahora depende de un regresor que no es $X_{0i} = 1$, sino que es $\Phi(\widehat{\alpha_\iota})$, y por tanto es como si no hubiera un coeficiente constante

- Los estimadores obtenidos por este método no requieren que el error se distribuya normal y/o que los errores sean homoscedásticos, pero si se cumplen ambas condiciones, se obtendrán estimadores menos eficientes de los que se obtendrían por máxima verosimilitud

Los modelos de dos ecuaciones

- Al usar el modelo Tobit, una de las suposiciones que se hacían es que los individuos con un valor de $Y_i = a$ (la solución de esquina) eran individuos en los cuales tenían un valor de $Y_i^* > a$ para la variable latente (dependiendo de la definición de la variable observada)

    - Lo que esto significa esta suposición es que el efecto de los regresores $\boldsymbol{x_i}$ en la probabilidad de que $Y_i^* > a$ es el mismo que en el valor esperado de la variable Y_i^* cuando $Y_i^* > a$ (dado que solo se tienen unos coeficientes $\boldsymbol{\beta}$ para ambas etapas)

        - Esta restricción es poco realista, dado que el efecto de las variables no siempre es el mismo para ambas etapas, pero el modelo considera que los coeficientes son los mismos

    - Para resolver este problema, se pueden plantear modelos en dos partes o *hurdle models* en los cuales se separan ambas etapas: una para la probabilidad de que $Y_i^* > a$, y otra para el valor esperado de Y_i^* cuando $Y_i^* > a$

        - Se separa la probabilidad de pertenecer a la distribución censurada $P(Y_i^* > a)$ del valor que se tiene dentro de esta distribución $E(Y_i^* | \boldsymbol{x_i}; Y_i^* > a)$

        - Aunque no se usen los mismos regresores ni el mismo efecto para cada decisión o parte (al tener diferentes coeficientes), puede haber correlación entre los errores, lo cual haría que las variables se determinaran de manera endógena

- Para poder lidiar con los regresores y la posible correlación entre términos de errores, se pueden plantear modelos aún más generales, como el modelo de Heckman, el *double hurdle model* de Cragg y el *two part model* (2P)

    - La ecuación de participación en los modelos depende de una variable latente d_i^* que determina una variable binaria D_i

$$d_i^* = \boldsymbol{z_i'\gamma} + u_i \quad and \quad D_i = \begin{cases} 1 \ \ if \ d_i^* > a \\ 0 \ \ if \ d_i^* \leq a \end{cases}$$

        - El término de error de la variable latente se distribuye de manera normal estándar (aunque se puede asumir una distribución normal para después estandarizar con la desviación típica, es normal usar la normal estándar)

$$u_i \sim N(0,1)$$

- La ecuación de intensidad en los modelos es general y sirve tanto para las observaciones censuradas como para las no censuradas, de modo que se modela una variable latente Y_i^*

$$Y_i^* = \boldsymbol{x_i'\beta} + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

  - El término de error en esta ecuación se distribuye de manera normal

- Existen dos mecanismos de observación: uno en el que se asume que $Y_i^* = a$ cuando $D_i = 0$ y/o $Y_i^* \leq a$, y otro en el que se asume que Y_i no se observa cuando $D_i = 0$

$$(Cragg\ \&\ 2P) \quad Y_i = Y_i^*\ if\ D_i = 1\ \&\ Y_i^* > a\ and\ Y_i = a\ otherwise$$

$$(Heckman) \quad Y_i = Y_i^*\ if\ D_i = 1\ \ and\ Y_i\ is\ unobs.\ if\ D_i = 0$$

  - El primer mecanismo de observación permite obtener el *double hurdle model* y el *two part model,* mientras que el segundo permite obtener el modelo de Heckman. El modelo de Heckman es un modelo de selección muestral porque permite estimar cuando hay un problema de selección muestral, mientras que el *double-hurdle* o el *two-part model* no lo son porque tienen en cuenta las observaciones $Y_i = a$

  - En el caso del modelo de dos partes y de Cragg, se asume que se observan los regresores $\boldsymbol{x_i}$ y que estos hacen que $Y_i = a$ (dado que se pone la condición para cualquier otra circunstancia) para los participantes. Esto hace que, si $\boldsymbol{x_i}$ se observa, las observaciones $Y_i = a$ puedan contribuir a la verosimilitud para la muestra completa y que haya dos fuentes para las cuales $Y_i = a$

  - No obstante, en el modelo de Cragg no se puede observar D_i, mientras que en el de dos partes sí. Esto hace que los valores $Y_i = a$ en el primer modelo puedan darse por no participar o como una solución de esquina (no se sabe exactamente al no observarse), pero que en el segundo modelo se pueda saber si proviene de uno o de otro

  - En el modelo de Heckman se asume que no se observan los regresores $\boldsymbol{X_i}$, por lo que no se puede observar una Y_i^* ni determinar una Y_i para $D_i = 0$. Esto hace que, si $\boldsymbol{X_i}$ no se observa, los individuos que no participan no contribuyan a la verosimilitud

- Los errores pueden estar correlacionados, por lo que se pueden determinar las variables de manera endógena y los errores seguirán una distribución normal bivariante con correlación ρ

$$(u_i, \varepsilon_i) \sim N(0,0,1,\sigma^2,\rho)$$

- La esperanza de Y_i para los individuos que están dentro de la muestra censurada permite obtener una expresión en la que se ve como se incluye la posibilidad de correlación y endogeneidad en los modelos

$$E(Y_i|\boldsymbol{x_i}, \boldsymbol{z_i}; D_i = 1) = \boldsymbol{x_i'\beta} + E(\varepsilon_i|\boldsymbol{x_i}, \boldsymbol{z_i}; D_i = 1)$$

$$\Rightarrow E(Y_i|\boldsymbol{x_i}, \boldsymbol{z_i}; D_i = 1) = \boldsymbol{x_i'\beta} + E(\varepsilon_i|\boldsymbol{x_i}, \boldsymbol{z_i}; u_i > a - \boldsymbol{z_i'\gamma})$$

$$\Rightarrow E(Y_i|\boldsymbol{x_i}, \boldsymbol{z_i}; D_i = 1) = \boldsymbol{x_i'\beta} + \rho\sigma\lambda(a - \boldsymbol{Z_i'\gamma})$$

$$\Rightarrow E(Y_i|\boldsymbol{x_i}, \boldsymbol{z_i}; D_i = 1) = \boldsymbol{x_i'\beta} + \rho\sigma\frac{\phi(a - \boldsymbol{z_i'\gamma})}{1 - \Phi(a - \boldsymbol{z_i'\gamma})}$$

$$\Rightarrow Y_i = \boldsymbol{x_i'\beta} + \rho\sigma\frac{\phi(a - \boldsymbol{z_i'\gamma})}{1 - \Phi(a - \boldsymbol{z_i'\gamma})} + \eta_i$$

  - El último término (la *ratio* inversa de Mills) puede variar de especificación dependiendo de la definición de la variable D_i, y se utiliza $\boldsymbol{z_i'\gamma}$ porque la variable ya se distribuye de manera normal estándar

  - Para el modelo de dos partes, se pueden utilizar formas funcionales tales como una log-normal o una truncada normal

- A partir de los efectos marginales, se puede comprobar que el signo del efecto marginal de un regresor (ya sea $\boldsymbol{x_i}$ o $\boldsymbol{z_i}$) depende del signo del coeficiente. Por lo tanto, se puede interpretar el signo del coeficiente, pero no la magnitud

  - No obstante, los coeficientes $\boldsymbol{\beta}$ son el efecto marginal latente (de la variable latente Y_i^*), por lo que son directamente interpretables. De este modo, se puede interpretar tanto la magnitud como el signo (dado que es una regresión lineal)

$$ME_k^{lat} = \frac{\partial E(Y_i^*|\boldsymbol{x_i})}{\partial X_{ki}} = \beta_k$$

  - Si $\boldsymbol{X_i}$ y $\boldsymbol{Z_i}$, y son diferentes, el efecto marginal de incrementar una unidad en el regresor X_{ik} es el coeficiente β_k, de modo que

el efecto marginal latente coincide con el efecto marginal de la variable dependiente observable. No obstante, $\boldsymbol{z_i}$ no afecta

$$ME_k = \frac{\partial E(Y|\boldsymbol{x_i}; D_i = 1)}{\partial X_{ki}} = \beta_k \quad only\ if\ \boldsymbol{x} \neq \boldsymbol{z}$$

- En el modelo de Heckman, el efecto en Y_i (siendo censurada) de un incremento unitario de una variable X_{ki} se debe interpretar como el efecto que tiene sobre aquellos individuos que tienen valores dentro de la distribución truncada, pero no para aquellos que la tienen fuera, dado que no se observan valores para $D_i = 0$

- En el *double hurdle model* y en el *two-part model*, el efecto en Y_i (siendo censurada) de un incremento unitario de una variable X_{ki} se debe interpretar como el efecto que tiene sobre aquellos individuos que tienen valores dentro de la distribución censurada y fuera

- A partir de la definición de los modelos, se puede ver como la función de densidad de la variable censurada Y_i es una combinación entre funciones continuas y discretas

  - Para el *double hurdle model* de Cragg, la función de verosimilitud no tendrá en cuenta $D_i = 0$ porque no se observa, de modo que solo tiene en cuenta las condiciones de las observaciones (que $D_i = 1$)

$$L = \prod_{Y_i = a} [1 - P(D_i = 1; Y_i^* > a)] \prod_{Y_i < a} P(D_i = 1; Y_i^* = Y_i)$$

$$1 - P(D_i = 1; Y_i^* > a) = 1 - P(u_i > a - \boldsymbol{z_i'\gamma}; \varepsilon_i > a - \boldsymbol{x_i'\beta}) =$$

$$= 1 - N(\boldsymbol{z_i'\gamma} - a, \boldsymbol{x_i'\beta} - a, 1, \sigma, \rho)$$

$$P(D_i = 1; Y_i^* = Y_i) = P(u_i > a - \boldsymbol{z_i'\gamma}; \varepsilon_i = Y_i - \boldsymbol{x_i'\beta}) =$$

$$= \int_{-\boldsymbol{z_i'\gamma}}^{\infty} N(u, \boldsymbol{x_i'\beta} - a, 1, \sigma, \rho)\, du =$$

$$= \frac{\phi\left(\frac{Y_i - \boldsymbol{x_i'\beta}}{\sigma}\right)}{\sigma}\left[1 - \Phi\left(\frac{-\boldsymbol{z_i'\gamma} - \rho\left(\frac{Y_i - \boldsymbol{x_i'\beta}}{\sigma}\right)}{\sqrt{1-\rho^2}}\right)\right]$$

- Para el modelo de dos partes, la función de verosimilitud si tendrá en cuenta $D_i = 0$ y por tanto lo puede utilizar para la estimación:

$$f(Y_i|\boldsymbol{z_i}) = \begin{cases} P(D_i = 1|\boldsymbol{z_i}) f(Y_i|\boldsymbol{x_i}; D_i = 1) & when\ Y_i > a \\ P(D_i = 0|\boldsymbol{z_i}) & when\ Y_i = a \end{cases}$$

$$f(Y_i|\boldsymbol{x_i}; D_i = 1) = density\ of\ the\ rand.var. when\ < a$$

$$P(D_i = 1\ or\ 0|\boldsymbol{z_i}) = probit\ or\ logit\ specification$$

- Para el modelo de Heckman, la función de verosimilitud será la siguiente (todo condicionado a los regresores correspondientes):

$$L = \prod_{Y_i = a} P(D_i = 0) \prod_{Y_i < a} P(D_i = 1; Y_i^* = Y_i)$$

$$P(D_i = 0) = P(u_i \leq a - \boldsymbol{z_i'\gamma}) = 1 - \Phi(a - \boldsymbol{z_i'\gamma})$$

$$P(D_i = 1; Y_i^* = Y_i) = P(u_i > a - \boldsymbol{z_i'\gamma}; \varepsilon_i = Y_i - \boldsymbol{x_i'\beta}) =$$

$$= \int_{-\boldsymbol{z_i'\gamma}}^{\infty} N(u, \boldsymbol{x_i'\beta} - a, 1, \sigma, \rho)\, du =$$

$$= \frac{\phi\left(\frac{Y_i - \boldsymbol{x_i'\beta}}{\sigma}\right)}{\sigma} \left[1 - \Phi\left(\frac{-\boldsymbol{z_i'\gamma} - \rho\left(\frac{Y_i - \boldsymbol{x_i'\beta}}{\sigma}\right)}{\sqrt{1 - \rho^2}}\right)\right]$$

- Cuando la correlación es nula, entonces se usa la probabilidad de dos variables independientes (se multiplican las probabilidades para cada variable y no se usa la distribución bivariante)

Los modelos de duración

- Un tipo de modelos en donde se utilizan variables censuradas son los modelos de duración. En estos modelos se considera el tiempo hasta que algún tipo de transición como la duración, y la transición misma como un evento
    - El rol que juega la censura en este tipo de modelos es que en casi todos los casos en los que se estudia datos de duración, algunos o muchos de los *spells* que se observan no acaban en transiciones
        - El término *spell* se refiere a las diferentes variables de duración que se pueden medir, y los datos de estos *spells* están, en efecto,

censurados, por lo que los modelos de duración tienen esto en cuenta explícitamente

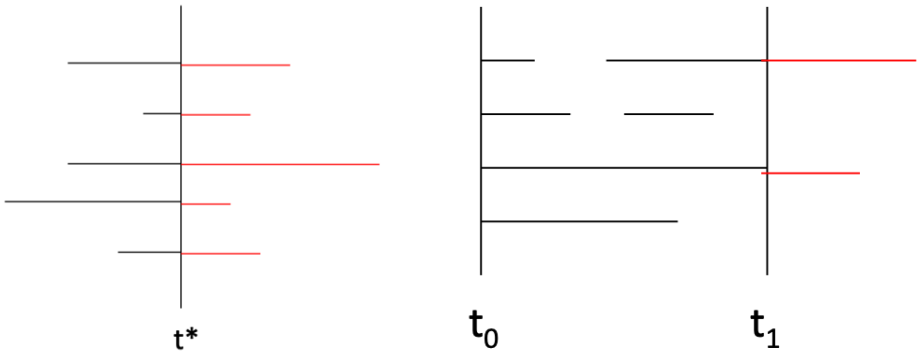
- Los aspectos más interesantes a analizar con estos modelos son aspectos relacionados con la probabilidad y el tiempo hasta que ocurre una transición

  - Aunque es interesante la duración de un evento, también es interesante la probabilidad de que un evento realice una transición en un periodo futuro condicional a la duración que ha tenido hasta ahora

  - El análisis de tiempo hasta el fallo o tiempo de supervivencia (*time until failure* o *survival time*), definido como el tiempo que un evento desde que comienza hasta que ocurre una transición, también es interesante de analizar

- La variable de interés es la duración de un evento desde que comienza hasta que termina o se mide, de modo que la medida puede preceder la terminación

  - Las observaciones de las duraciones suelen tener una forma de datos transversales o *cross-sectional data*, o con una forma de datos longitudinales o *longitudinal data*



  - Como se puede ver, la duración es el valor de la variable desde que comienza el proceso (puede haber diferentes comienzos para diferentes individuos) hasta t^* para datos transversales, mientras que la duración es el valor de la variable desde t_0 hasta t_1 (hay un solo comienzo para cada individuo

  - Normalmente se representa el largo de estos *spells* con una variable T, mientras que la duración realizada (observada en los datos) se representa con t

  - La censura es un problema general que comúnmente es debido a que se suele medir cuando el proceso aún sigue, de modo que las

observaciones de individuos cuyo evento sigue en marcha están censuradas necesariamente

- En consecuencia, la duración o supervivencia de los individuos en una muestra para datos transversales es como mínimo $T \geq t^*$, pero no igual a t^* (dado que no se suele observar cuando hay una transición), por lo que se considera que todos los *spells* están incompletos y censurados. Para los datos longitudinales, en cambio, pueden haber *spells* completos e incompletos, de modo que no hay una duración mínima

- La estimación de este tipo de modelos tiene que tener en cuenta esta naturaleza censurada de los datos por el efecto de esta censura en la media y la varianza y en la inferencia (sesgo por omisión de variables)

- Representando la duración (el *spell*) como una variable aleatoria T, hay aspectos de su distribución de probabilidad que permitirán estudiar y hacer un análisis regresivo a partir de modelos

  - Suponiendo que la variable T tiene una función de densidad $f(t)$, se puede obtener la función de distribución de probabilidad acumulada y la función de supervivencia

    - Integrando la función de densidad de la variable de 0 a t, se puede obtener la función de probabilidad acumulada

$$F(t) = \int_0^t f(s)\, ds = P(T \leq t)$$

    - La función de supervivencia o *survival function* se define como la probabilidad de que un *spell* sea de una duración de al menos t, por lo que es la probabilidad a la derecha de t

$$S(t) = 1 - \int_0^t f(s)\, ds = 1 - P(T \leq t) = P(T \geq t)$$

  - Para poder saber cuál es la probabilidad de que un *spell* finalice en un momento futuro, se utiliza la función de distribución y la noción de tasa de riesgo

    - La probabilidad de que un *spell* finalice en un momento futuro $t + \Delta t$ dado que ha durado hasta el momento t (como mínimo) es la siguiente:

$$l(t, t + \Delta t) = P(t \leq T \leq t + \Delta t \mid T \geq t) = \frac{F(t + \Delta t) - F(t)}{S(t)}$$

- Una función útil para caracterizar este aspecto de la distribución de probabilidad es la función de riesgo o *hazard function*, que se entiende como la velocidad a la que los *spells* sufren una transición después de durar t, dado que duran como mínimo t. Se puede demostrar, además, que la tasa es la semielasticidad negativa de la función de supervivencia, por lo que la función de densidad es el producto entre $S(t)$ y $\lambda(t)$

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t S(t)} = \frac{f(t)}{S(t)}$$

$$\lambda(t) = -\frac{d \ln S(t)}{dt} = -\frac{d \ln S(t)}{dS(t)} \frac{dS(t)}{t} = -\frac{-f(t)}{S(t)} = \frac{f(t)}{S(t)}$$

$$\Rightarrow f(t) = \lambda(t) S(t)$$

- Una función útil que también se usa en los modelos es la función de riesgo integrada, la cual permite obtener una expresión alternativa para la función de supervivencia

$$\Lambda(t) = \int_0^t \lambda(s)\, ds \;\Rightarrow\; \Lambda(t) = \int_0^t -\frac{d \ln S(t)}{dt}\, ds = -\ln S(t)$$

$$\Rightarrow S(t) = e^{-\Lambda(t)}$$

- Además, la derivada de la función de riesgo permite saber cómo los *spells* dependen de la duración t, de modo que, si es positiva, más duración realizada (una t mayor) significa que la probabilidad de que finalice el *spell* es mayor, mientras que una negativa significa lo opuesto

$$\frac{d\lambda(t)}{dt} > 0 \;\; (positive\ dependence)$$

$$\frac{d\lambda(t)}{dt} < 0 \;\; (negative\ dependence)$$

- Dependiendo de si los datos son transversales o longitudinales, la función de máxima verosimilitud varía

  - Para los datos transversales, la función de máxima verosimilitud y su logaritmo vienen dadas por las siguientes expresiones:

$$L = \prod_{i=1}^{N} \frac{S(t)}{E(T)} = \prod_{i=1}^{N_1} \frac{1 - F(t)}{\int_0^\infty [1 - F(t)]\, dt}$$

$$\Rightarrow \ln L = \sum_{i=1}^{N_1} \ln S(t) + \sum_{i=N_1+1}^{N} \ln E(t)$$

- Para los datos longitudinales, la función de máxima verosimilitud y su logaritmo vienen dadas por las siguientes expresiones:

$$L = \prod_{i=1}^{N_1} P(T_i = t_i) \prod_{i=N_1+1}^{N} P(T_i > t_i) = \prod_{i=1}^{N_1} f(t) \prod_{i=N_1+1}^{N} S(t)$$

$$\Rightarrow \ln L = \sum_{i=1}^{N_1} \ln f(t) + \sum_{i=N_1+1}^{N} \ln S(t)$$

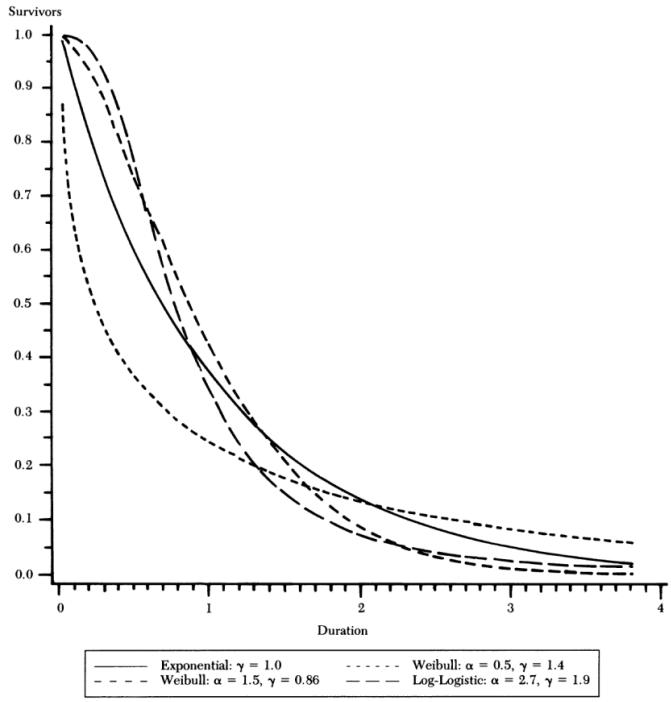
$where\ 1, \dots, N_1\ \ are\ completed\ and\ N_1 + 1, \dots, N\ are\ incompleted\ spells$

- Representando la duración (el *spell*) como una variable aleatoria T, se puede aplicar un análisis regresivo para la muestra de *spells* observados, por lo que se puede caracterizar la duración esperada condicionada a unos regresores y la distribución de probabilidad

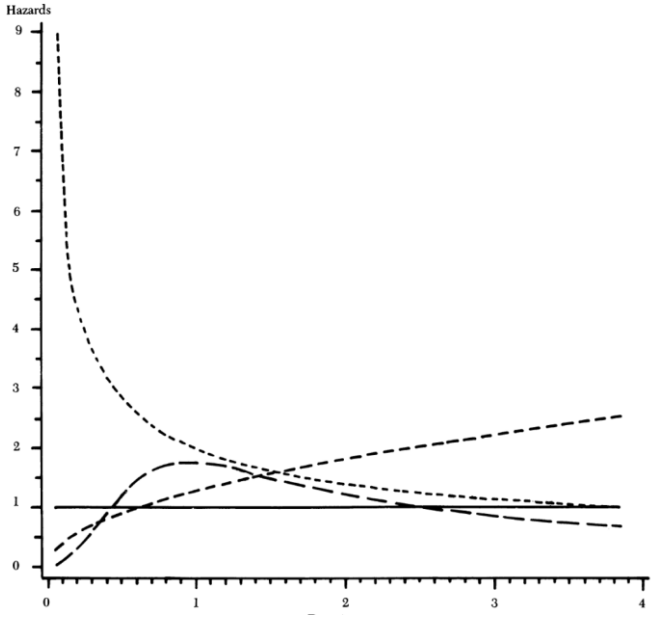
  - La función de riesgo es más útil en el análisis que no la función de supervivencia o la densidad, por lo que es mejor modelar la función de riesgo

    - Como el interés está en si hay más o menos probabilidad de que un *spell* termine en el momento t dado que ha durado hasta el momento t, la dependencia de la duración es lo más relevante y eso implica que también lo es la función de riesgo

    - La elección del modelo dependerá de cómo se ajusten los datos a cada uno de ellos. Mirando la función de supervivencia, se puede ver que los modelos principales son muy similares, y, por tanto, no es muy útil modelar esta función



- No obstante, la función de riesgo para cada modelo exhibe un comportamiento diferenciado, de modo que es útil modelar esta función con tal de ajustarse a los datos



- Los modelos más importantes y más usados para el análisis de regresión son el modelo exponencial, el modelo Weibull, el modelo log-normal y el modelo log-logístico

Distribution	*Hazard Function, $\lambda(t)$*	*Survival Function, S(t)*
Exponential	λ,	$S(t) = e^{-\lambda t}$
Weibull	$\lambda p(\lambda t)^{p-1}$,	$S(t) = e^{-(\lambda t)^p}$
Lognormal	$f(t) = (p/t)\phi[p\ln(\lambda t)]$	$S(t) = \Phi[-p\ln(\lambda t)]$
	[$\ln t$ is normally distributed with mean $-\ln\lambda$ and standard deviation $1/p$.]	
Loglogistic	$\lambda(t) = \lambda p(\lambda t)^{p-1}/[1+(\lambda t)^p]$,	$S(t) = 1/[1+(\lambda t)^p]$
	[$\ln t$ has a logistic distribution with mean $-\ln\lambda$ and variance $\pi^2/(3p^2)$.]	

- El modelo exponencial es un modelo en el que se asume que la función de riesgo es constante, por lo que la dependencia de la duración es nula

  - En este caso, se asume que la función de riesgo es una constante λ, lo cual implica que la función integrada será una función del tiempo y que la dependencia con la duración es nula, por lo que la probabilidad no se altera por la duración

$$\lambda(t) = \gamma \qquad \Lambda(t) = \gamma t \qquad \frac{d\lambda(t)}{dt} = 0$$

  - A partir de la función de riesgo, se puede ver que la función de supervivencia y las funciones de distribución y densidad son las siguientes:

$$S(t) = e^{-\gamma t} \qquad F(t) = 1 - e^{-\gamma t} \qquad f(t) = \lambda e^{-\gamma t}$$

  - Adicionalmente, se puede demostrar que los momentos de la distribución de probabilidad de T son los siguientes:

$$E(T) = \gamma^{-1} \qquad Var(T) = \gamma^{-2}$$

- El modelo Weibull es un modelo similar al exponencial, pero en el que la función de riesgo no es constante, haciendo que exista una dependencia de la duración no nula

  - En este caso, se asume que la función de riesgo depende de la duración realizada o tiempo t y de un parámetro α

$$\lambda(t) = \gamma\alpha t^{\alpha-1} \qquad \Lambda(t) = \gamma t^{\alpha} \qquad \frac{d\lambda(t)}{dt} = \gamma\alpha(\alpha-1)t^{\alpha-2}$$

  - La dependencia de la duración dependerá de estos parámetros. Si $\alpha > 1$, la dependencia es positiva y la probabilidad de fallida aumenta cuanto más haya durado el *spell*, mientras que se da lo opuesto si $\alpha < 1$. Si $\alpha = 1$, entonces el modelo es equivalente al exponencial y no hay dependencia de la duración

$$\frac{d\lambda(t)}{dt} = \gamma\alpha(\alpha-1)t^{\alpha-2} \Rightarrow \begin{cases} \frac{d\lambda(t)}{dt} > 0 \ \ if \ \alpha > 1 \\ \frac{d\lambda(t)}{dt} = 0 \ \ if \ \alpha = 1 \\ \frac{d\lambda(t)}{dt} < 0 \ \ if \ \alpha < 1 \end{cases}$$

- A partir de la función de riesgo, se puede ver que la función de supervivencia y las funciones de distribución y densidad son las siguientes:

$$S(t) = e^{-\gamma t^{\alpha}} \qquad F(t) = 1 - e^{-\gamma t^{\alpha}} \qquad f(t) = \gamma\alpha t^{\alpha-1} e^{-\gamma t^{\alpha}}$$

- Adicionalmente, se puede demostrar que la esperanza de la distribución de probabilidad de T es la siguiente:

$$E(T) = \gamma^{-\frac{1}{\alpha}} \Gamma\left(1 + \frac{1}{\alpha}\right)$$

- El modelo log-logístico es un modelo similar al Weibull, pero en el que la integral de la función de riesgo es un logaritmo, lo cual provoca que la dependencia de la duración no sea constante y pueda variar con la duración

  - En este caso, se asume que la función de riesgo depende de la duración realizada o tiempo t y de un parámetro α

$$\lambda(t) = \frac{\gamma\alpha t^{\alpha-1}}{1+\gamma t^{\alpha}} \qquad \Lambda(t) = \ln(1+\gamma t^{\alpha})$$

  - La dependencia de la duración dependerá de estos parámetros. Si $\alpha > 1$, la dependencia es positiva en un principio y después es negativa, haciendo que la probabilidad de fallida aumente y después se reduzca, mientras que la dependencia será negativa si $\alpha \leq 1$. La duración t^{max} en la que cambia el signo de la dependencia se puede encontrar a partir de su expresión

$$\frac{d\lambda(t)}{dt} = \frac{\gamma\alpha(\alpha-1)t^{\alpha-2}}{1+\gamma t^{\alpha}} - \frac{(\gamma\alpha t^{\alpha-1})^2}{(1+\gamma t^{\alpha})^2}$$

$$\Rightarrow \begin{cases} \frac{d\lambda(t)}{dt} > 0 \ \ and \ \ \frac{d\lambda(t)}{dt} < 0 \ \ after \ \ if \ \alpha > 1 \\ \frac{d\lambda(t)}{dt} < 0 \ \ if \ \alpha \leq 1 \end{cases}$$

$$\frac{d\lambda(t)}{dt} = 0 \ \ \Rightarrow \ t^{max} = \left(\frac{\alpha-1}{\gamma}\right)^{\frac{1}{\alpha}}$$

- A partir de la función de riesgo, se puede ver que la función de supervivencia y las funciones de distribución y densidad son las siguientes:

$$S(t) = \frac{1}{1+\gamma t^{\alpha}} \qquad F(t) = 1 - \frac{1}{1+\gamma t^{\alpha}} \qquad f(t) = \frac{\gamma \alpha t^{\alpha-1}}{(1+\gamma t^{\alpha})^2}$$

- Adicionalmente, se puede demostrar que la esperanza de la distribución de probabilidad de T es la siguiente:

$$E(T) = \frac{\gamma^{-1}}{\sin\left(\frac{\pi}{\alpha}\right)}$$

- Dos limitaciones de los modelos anteriores es que no hay lugar para factores externos y que, aunque se incluyan, puede haber diferencias individuales en las distribuciones remanentes por una especificación incompleta. Estas se pueden superar a través de añadir regresores y de tener en cuenta esta heterogeneidad

  - La adición de regresores a los modelos de duración es bastante simple, aunque la interpretación de los coeficientes no lo es tanto. Parar introducir la heterogeneidad, se asume que el parámetro γ es una función que depende de $\boldsymbol{x_i}$

$$\gamma_i = e^{\boldsymbol{x_i'\beta}}$$

    - Normalmente se asume que estos regresores $\boldsymbol{x_i}$ tienen un valor constante. Para datos transversales, se puede asumir que se escoge el valor de $\boldsymbol{x_i}$ en t^*, mientras que, para datos longitudinales, se puede asumir que los regresores toman un valor fijo de t_0 a t_1

    - Además, se suele asumir que la duración realizada t sigue una distribución normal, aunque esta suposición no suele ser la mejor y se pueden asumir otro tipo de distribuciones

    - En este caso, la función de riesgo no depende del tiempo porque se asume que los regresores son constantes, pero sí depende del individuo

    - Como hacer que λ_i sea una función de sus regresores es equivalente a realizar un cambio de unidades en el eje del tiempo o de las ordenadas, estos modelos se suelen llamar modelos de duración acelerada

- El efecto de un cambio en X_{ki} a partir de un modelo exponencial o Weibull permite ver como el signo es el mismo que el del coeficiente, y se suele expresar a través de la semielasticidad para una mejor interpretación

$$(exponential) \quad \frac{d\lambda_i}{dX_{ki}} = \beta_k e^{\boldsymbol{x_i'\beta}} \;\Rightarrow\; \frac{d\ln\lambda_i}{dX_{ki}} = \beta_k$$

$$(Weibull) \quad \frac{d\lambda_i}{dX_{ki}} = \beta_k e^{\boldsymbol{x_i'\beta}} \alpha t^{\alpha-1} \;\Rightarrow\; \frac{d\ln\lambda_i}{dX_{ki}} = \beta_k$$

  - Dado que para el modelo exponencial y Weibull la semielasticidad es el coeficiente del regresor, se puede interpretar el efecto de aumentar un regresor como en una regresión con logaritmos en la variable dependiente

  - El efecto de un incremento marginal en un regresor para el modelo log-logístico y la semielasticidad de este es más complejo que para los otros modelos

$$\frac{d\lambda_i}{dX_{ki}} = \beta_k e^{\boldsymbol{X_i'\beta}} \frac{\alpha t^{\alpha-1}}{\left(1 + e^{\boldsymbol{x_i'\beta}} t^{\alpha}\right)^2} \;\Rightarrow\; \frac{d\ln\lambda_i}{dX_{ki}} = \frac{\beta_k}{1 + e^{\boldsymbol{x_i'\beta}} t^{\alpha}}$$

$$\frac{d\lambda_i}{dX_{ki}} = 0 \;\Rightarrow\; \beta_k e^{\boldsymbol{x_i'\beta}} \frac{\alpha t^{\alpha-1}}{\left(1 + e^{\boldsymbol{x_i'\beta}} t^{\alpha}\right)^2} = 0 \Rightarrow \beta_k e^{\boldsymbol{x_i'\beta}} \alpha t^{\alpha-1}$$

  - En este caso, el signo de la dependencia será el mismo al de los coeficientes, de modo que aumentar en una unidad un regresor X_{ki} hace que la probabilidad de que el *spell* finalice en un momento futuro aumente, y viceversa

- Otra forma de expresar el efecto en la probabilidad de fallida de los regresores es a través de la función de riesgo proporcional, definida como la *ratio* de funciones de riesgo

$$\frac{\lambda_i}{\lambda_j} = \frac{e^{\boldsymbol{x_i'\beta}}}{e^{\boldsymbol{x_j'\beta}}} = e^{(\boldsymbol{x_i - x_j})'\boldsymbol{\beta}}$$

  - El efecto de aumentar una unidad en un regresor X_{ki} es la reducción en λ_i con respecto a λ_j

  - Tanto el modelo exponencial como el Weibull tienen una función de riesgo proporcional, mientras que el modelo log-logístico no. Esto hace que, para los dos primeros modelos, se pueda interpretar el exponencial del coeficiente $e^{\boldsymbol{\beta}}$ como el

cambio en λ al variar una unidad de un regresor (como *log-level*). No obstante, esto no aplica al modelo log-logístico

- Para lidiar con la heterogeneidad que no se observa a partir de los datos individuales, se suele incluir una variable aleatoria v_i que siga una distribución en la función de riesgo multiplicando

  - De este modo, esta variable aporta aleatoriedad que no es observada en la muestra en la función de riesgo. No obstante, su introducción hace que a veces se den resultados contradictorios al sentido común

  - Asumiendo un modelo Weibull, se puede ver el efecto en la función de supervivencia

$$\lambda_i(t|v_i) = \gamma_i \alpha t^{\alpha-1} v_i \;\Rightarrow\; S_i(t|v_i) = e^{-\int_0^\infty \lambda_i(s|v_i)\, ds} = e^{-\gamma_i t^\alpha v_i}$$

  - La distribución comúnmente asumida es una distribución gamma con umbral θ y factor de forma 1

$$v_i \sim \Gamma(1, \theta)$$

- Los modelos paramétricos pueden distorsionar las funciones de riesgo estimadas, por lo que se podría hacer una representación menos restrictiva

  - El estimador de la función de supervivencia para la muestra cuando no hay censura es la siguiente:

$$\hat{S}(t) = \frac{n^{\underline{o}}\ of\ spells\ such\ that\ T \geq t}{n^{\underline{o}}\ of\ total\ spells}$$

    - Esto se da porque, al no haber censura, entonces el mejor estimador para la probabilidad de que la duración sea mayor a t es la proporción de los spells (observaciones) que son mayores a t dentro de la muestra

  - El estimador producto límite de Kaplan-Meier es un estimador estrictamente empírico y no paramétrico para la estimación de la función de riesgo y de supervivencia

$$\hat{\lambda}(t) = \frac{h_j}{n_j} \qquad \hat{S}(t_j) = \prod_{i=1}^{j} \frac{n_i - h_i}{n_i} = \prod_{i=1}^{j} 1 - \hat{\lambda}(t_i)$$

    - Se asume que hay K duraciones completas que corresponden a *spells* concretos están ordenadas en orden ascendente, y que $K < N$ porque hay *spells* censurados y vínculos

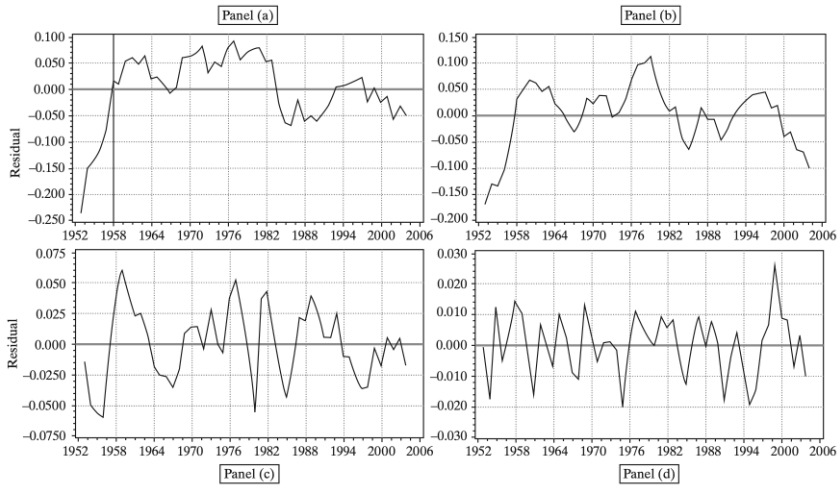
$$t_1 < t_2 < t_3 < \cdots < t_K$$

- En este caso, h_j representa el número de observaciones o duraciones completas con una duración t_j, m_j representa el número de observaciones censuradas entre t_j y t_{j+1}, m_K representa el número de observaciones con una duración mayor a t_K y n_j representa el número de *spells* que no se han censurado ni completado antes de t_j

$$n_j = \sum_{i \geq j}^{K} (m_i + h_i)$$

La correlación serial

- Las series temporales normalmente presentan autocorrelación o correlación serial en las perturbaciones cuando se miran diferentes periodos
  - Una explicación para la autocorrelación es que los factores relevantes omitidos de la regresión de series temporales, como aquellos incluidos, están correlacionados en el tiempo
    - Este hecho se puede deber a que la correlación serial de los factores tendría que haberse incluido en la regresión. Por lo tanto, puede haber autocorrelación inducida por una mala especificación del modelo



  - Los problemas para la estimación y la inferencia causados para la autocorrelación son parecidos (aunque más complejos) a aquellos causados por la heteroscedasticidad

        - Los estimadores de mínimos cuadrados ordinarios son ineficientes y la inferencia basada en estos estimadores se ve afectada adversamente

        - Dependiendo del proceso subyacente, los estimadores de mínimos cuadrados generalizados y los de mínimos cuadrados generalizados factibles se pueden usar para solucionar este problema

- Un modelo de series temporales normalmente describe el camino de una variable y en términos de factores $\boldsymbol{x}_t$ contemporáneos, las perturbaciones ε_t (llamadas innovaciones) y sus valores pasados $y_{t-1}, y_{t-2}, \ldots$

    - Una serie temporal es una sola ocurrencia de un evento aleatorio (de la variable y en el momento t), de modo que el historial entero sobre el periodo examinado en las series temporales constituye una sola realización del proceso

        - En relación a economía, el proceso no se puede repetir, y no hay contraparte al muestreo con repetición con datos transversales o una replicación del experimento que involucre una serie temporal en física o ingeniería

        - Sin embargo, podría haberse dado una realización completamente diferente de toda la serie temporal

    - La secuencia de observaciones $\{y_t\}_{t=-\infty}^{t=\infty}$ es un proceso de serie temporal, el cual se caracteriza por su orden temporal y su correlación sistemática entre las observaciones de esta secuencia

        - La característica más importante de un proceso de series temporales es que, empíricamente, el mecanismo de generación de datos produce exactamente una realización de esta secuencia

        - Los resultados estadísticos basados en características muestrales no tienen relación con el muestreo aleatorio de una población, sino del muestreo aleatorio de las distribuciones de estadísticos construidos a partir de conjuntos de observaciones tomadas de esta realización en una ventana temporal $t = 1,2, \ldots, T$

        - En consecuencia, la teoría asintótica en este contexto está relacionada con el comportamiento de los estadísticos construidos de una muestra con un tamaño incremental en esta secuencia

    - Las propiedades de y_t como variable aleatoria en datos transversales son directas y resumidas convenientemente en una proposición sobre

su media y su varianza o de la distribución generadora de y_t. En este caso, eso es menos obvio

- Es común asumir que las innovaciones se generan independientemente de un periodo al siguiente y que la media es nula y la varianza es homoscedástica. Bajo estas suposiciones, la distribución de ε_t sería débilmente estacionaria

$$E(\varepsilon_t|\boldsymbol{X}) = E(\varepsilon_t) = 0 \qquad Var(\varepsilon_t|\boldsymbol{X}) = Var(\varepsilon_t) = \sigma_\varepsilon^2$$

$$Cov(\varepsilon_t, \varepsilon_s) = 0 \quad for \quad t \neq s$$

- Bajo las suposiciones, aunque la noción de muestreo aleatorio se tiene que extender a las series temporales ε_t, los resultados matemáticos basados en la noción aplican aquí

- No obstante, hay una diferencia obvia entre las series temporales de y_t y de ε_t: las observaciones de y_t en diferentes puntos en el tiempo están necesariamente correlacionadas

  - Suponiendo que la serie y_t es débilmente estacionaria y un modelo $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$, entonces se puede ver como la varianza de y_t es homoscedástica siempre que se cumpla que $|\beta_1| < 1$ (para que la varianza sea positiva y finita)

$$E(y_t) = \beta_0 + \beta_1 E(y_{t-1}) + E(\varepsilon_t) \;\Rightarrow\; E(y_t) = \frac{\beta_0}{1-\beta_1}$$

$$Var(y_t) = \beta_1^2 Var(y_{t-1}) + Var(\varepsilon_t) \;\Rightarrow\; Var(y_t) = \frac{\sigma_\varepsilon^2}{1-\beta_1^2}$$

  - Alternativamente, se puede visualizar sabiendo que la realización en un momento pasado depende de un momento anterior a ese, por lo que se puede expresar y_t como una suma infinita que depende de los coeficientes (como un proceso que genra los datos en un pasado infinito). Por lo tanto, estos tienen que ser $|\beta_1| < 1$ para que la suma no diverja

$$y_t = \beta_0 + \beta_1(\beta_0 + \beta_1(\beta_0 + \beta_1 y_{t-3} + \varepsilon_t) + \varepsilon_t) + \varepsilon_t$$

$$\Rightarrow\; y_t = \sum_{i=0}^{\infty} \beta_1^i(\beta_0 + \varepsilon_{t-i})$$

  - También se podría llegar a la misma conclusiones que con los últimos dos ejemplos si se asumen condiciones iniciales para y_0 y ε_0, de modo que se asume que la observación de esta serie temporal se inicia en un momento $t = 0$ en el que el proceso

subyacente ha alcanzado un estado en donde la distribución de la variable ya no cambia

- Excepto en casos muy especiales, se espera que todos los elementos en el vector aleatorio $y_t = (y_1, y_2, \dots, y_T)$ estén correlacionados (llamada autocorrelación), por lo que no se pueden utilizar los métodos de estimación para observaciones no correlacionadas vistos anteriormente no funcionan

  - Hay una contraparte a la estimación paramétrica para datos transversales, pero bajo suposiciones que hagan que los parámetros (en el sentido familiar) existan

  - Aún con estacionariedad, los resultados de muestra finita vistos anteriormente no son útiles

- Las series temporales pueden exhibir una variedad de patrones, por lo que ayuda mucho descomponer estas series en varios componentes

  - Al describir patrones de series temporales, se suelen utilizar tres conceptos básicos: la tendencia, la estacionalidad y la ciclicidad

    - La tendencia se puede entender como un crecimiento o decrecimiento prolongado en los datos

    - La estacionalidad se puede entender como un patrón que ocurre debido a factores estacionales como el momento del año o el dia de la semana. Esta siempre tiene una frecuencia fija y conocida

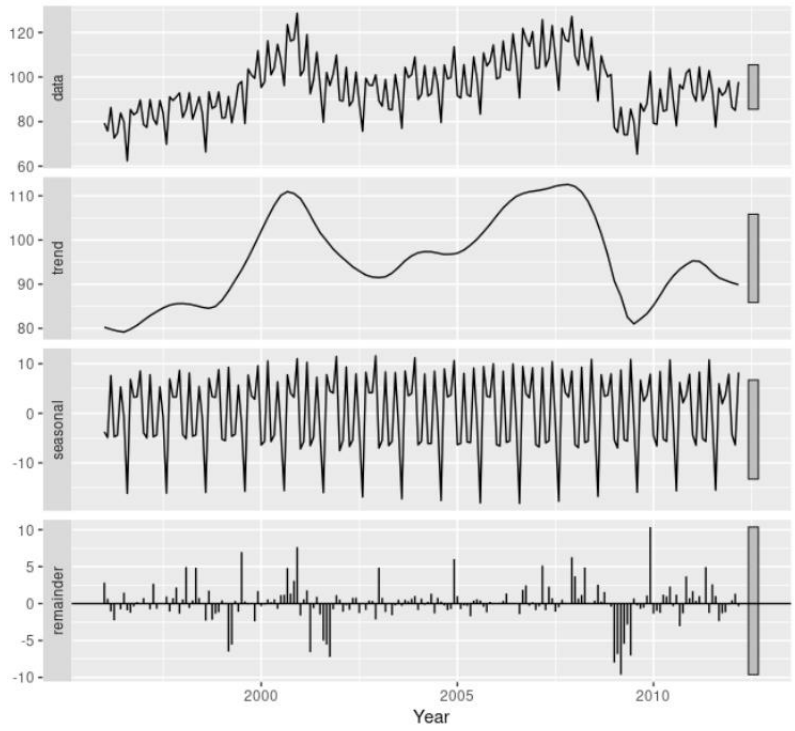
    - La ciclicidad se puede entender como las caídas y subidas que no son de una frecuencia fija. Estas se suelen dar por factores económicos y por el ciclo comercial

  - A partir de estos tres tipos de patrones, se proponen dos tipos de descomposiciones para los datos: la descomposición aditiva y la descomposición multiplicativa

$$y_t = T_t + S_t + C_t + R_t \quad (additive\ dec.)$$

$$y_t = T_t S_t C_t R_t \quad (multiplicative\ dec.)$$

    - En este caso, T_t representa el componente de tendencia, S_t representa el componente de estacionalidad, y R_t representa el componente residual, que son las fluctuaciones aleatorias o sistémicas. Debido a que la ciclicidad es muy difícil de identificar, se suele omitir en los paquetes más básicos de *software*



- La descomposición aditiva es más útil si la magnitud de las fluctuaciones del componente estacional no varia con el nivel de la serie temporal (la tendencia no es proporcional). Si, en cambio, la variación en el componente estacional es proporcional al componente tendencial, entonces la descomposición multiplicativa es más apropiada

- Una manera de utilizar la descomposición multiplicativa es hacer una transformación logarítmica de los datos hasta que la variación parezca ser estable en el tiempo, y así aplicar la descomposición aditiva (ya que serán descomposiciones equivalentes)

$$y_t = T_t S_t C_t R_t \;\Rightarrow\; \log y_t = \log T_t + \log S_t + \log C_t + \log R_t$$

- Ignorando el componente estacional, es posible hacer una estimación de la tendencia de una serie temporal a través de una regresión de mínimos cuadrados para el tiempo (el número de observaciones)

$$y_t = a + bt + \varepsilon_t$$

- Obviamente, la tendencia no tiene por qué ser perfectamente lineal, por lo que se puede generalizar a un modelo polinómico que se puede estimar por mínimos cuadrados ordinarios

$$y_t = a + b_1 t + b_2 t^2 + \cdots + b_p t^p + \varepsilon_t$$

- Aunque la tendencia sea un predictor natural de los valores de y_t, debido a que las innovaciones ε_t están autocorrelacionadas, es posible obtener mejores estimadores para y_t teniendo en cuenta este hecho

- Una vez vistas las nociones básicas sobre series temporales, se pueden obtener resultados teóricos utilizando conclusiones sobre la autocorrelación y proponer modelos para el análisis de series temporales

  - En las series temporales comunes, las perturbaciones son homoscedásticas, pero correlacionadas entre observaciones. Por lo tanto, siendo $\sigma^2\mathbf{\Omega}$ una matriz completa definida positiva con constante $Var(\varepsilon_t|\boldsymbol{X}) = \sigma^2$ en la diagonal, se obtiene el siguiente resultado:

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}) = \sigma^2\mathbf{\Omega}$$

    - Se debe asumir que $\mathbf{\Omega}_{t,s}$ es una función de $|t-s|$ (lo cual es una suposición de estacionariedad), pero no de t o de s únicamente. Esto implica que la covarianza entre observaciones t y s son una función solo de $|t-s|$, la distancia temporal entre las observaciones

    - Debido a que σ^2 no está restringida, se normaliza $\mathbf{\Omega}_{t,t} = 1$

  - A partir de esto, se puede definir la noción de autocovarianza y la de autocorrelación

    - La autocovarianza se define como la covarianza condicional de los errores a $\boldsymbol{X}$. La autocovarianza de dos errores en el mismo momento será la varianza de ε_t

$$\gamma_s \equiv Cov(\varepsilon_t, \varepsilon_{t-s}|\boldsymbol{X}) = Cov(\varepsilon_{t+s}, \varepsilon_t|\boldsymbol{X}) = \sigma^2\mathbf{\Omega}_{t,t-s}$$

$$\gamma_0 \equiv Cov(\varepsilon_t, \varepsilon_t|\boldsymbol{X}) = \sigma^2\mathbf{\Omega}_{t,t} = Var(\varepsilon_t|\boldsymbol{X}) = \sigma^2$$

    - La autocorrelación se define como la correlación condicional de los errores a $\boldsymbol{X}$

$$\rho_s \equiv Corr(\varepsilon_t, \varepsilon_{t-s}|\boldsymbol{X}) = \frac{Cov(\varepsilon_t, \varepsilon_{t-s}|\boldsymbol{X})}{\sqrt{Var(\varepsilon_t|\boldsymbol{X})}\sqrt{Var(\varepsilon_{t-s}|\boldsymbol{X})}} = \frac{\gamma_s}{\gamma_0}$$

    - Por lo tanto, se puede escribir la covarianza en términos de la esperanza, resultando en una matriz $\mathbf{\Gamma}$ de autocovarianzas, que es igual a la matriz de autocorrelaciones $\boldsymbol{R}$ multiplicada por γ_0

$$E(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2\mathbf{\Omega}_{t,t-s} = \mathbf{\Gamma} = \gamma_0\boldsymbol{R}$$

- Diferentes procesos implican diferentes patrones para la matriz $\boldsymbol{R}$, pero principalmente se pueden identificar tres procesos: el proceso autorregresivo, el proceso de media móvil y el proceso autorregresivo con media móvil

    - El caso más estudiado para los errores es el modelo autorregresivo de primer orden $AR(1)$, en donde u_t es estacionario no autocorrelacionado (es un ruido blanco) y ρ es un parámetro

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

    - Los procesos autorregresivos $AR(p)$ de mayor orden implican unos patrones más complejos, incluyendo, para algunos valores de los parámetros, un comportamiento cíclico en el comportamiento de las autocorrelaciones

$$\varepsilon_t = \gamma_1\varepsilon_{t-1} + \gamma_2\varepsilon_{t-2} + \cdots + \gamma_p\varepsilon_{t-p} + u_t$$

$$C(L)\varepsilon_t = u_t \quad for \quad C(L)x_t \equiv \left(1 - \gamma_1 L - \gamma_2 L^2 - \cdots - \gamma_p L^p\right)x_t$$

    - Las autorregresiones estacionarias están estructuradas de tal manera que la influencia de una perturbación dada se desvanece a medida que retrocede al pasado más distante, pero solo desaparece asintóticamente

    - En contraste, los procesos de media móvil tienen una memoria corta, y el proceso de media móvil más utilizado es el proceso $MA(1)$. Para este modelo, la memoria del proceso es de solo un periodo, lo cual se puede ver calculando las autocovarianzas

$$\varepsilon_t = u_t - \theta_1 u_{t-1}$$

$$\gamma_0 = Var(\varepsilon_t, \varepsilon_t) = \sigma_u^2(1 + \lambda_1^2)$$

$$\gamma_1 = Cov(\varepsilon_{t+1}, \varepsilon_t) = Cov(-\lambda_1 u_t, u_t) = -\lambda_1\sigma_u^2$$

$$\gamma_{s>1} = Cov(\varepsilon_{t+s}, \varepsilon_t) = 0$$

    - Los procesos de media móvil $MA(p)$ de mayor orden implican unos patrones más complejos, y estos se pueden expresar de la siguiente manera:

$$\varepsilon_t = u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2} - \cdots - \theta_q u_{t-q}$$

$$\varepsilon_t = D(L)u_t \quad for \quad D(L)x_t \equiv \left(1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_q L^q\right)x_t$$

- Este modelo también se puede expresar con coeficientes con signo positivo, aunque eso no cambia las propiedades teóricas del modelo (solo el signo)

$$\varepsilon_t = u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} - \cdots + \theta_q u_{t-q}$$

$$\varepsilon_t = -D(L)u_t \quad for \quad D(L)x_t \equiv \left(1 + \theta_1 L + \cdots - \theta_q L^q\right)x_t$$

- Finalmente, también es posible combinar los dos modelos mostrados anteriormente en un modelo autorregresivo con media móvil, denotado por $ARMA(p, q)$

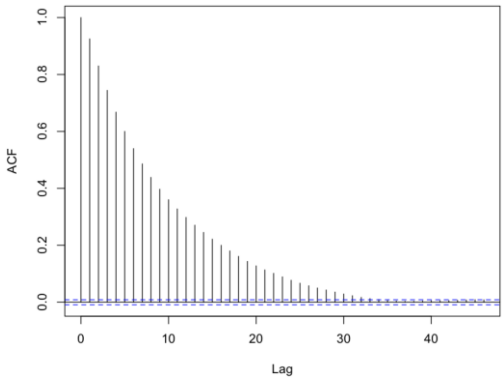
$$\varepsilon_t = \gamma_1 \varepsilon_{t-1} + \cdots + \gamma_p \varepsilon_{t-p} + u_t - \theta_1 u_{t-1} - \cdots - \theta_q u_{t-q}$$

$$C(L)\varepsilon_t = D(L)u_t$$

- Para poder seleccionar el orden de estos modelos, se necesita tener en cuenta la función de autocorrelación y la función de autocorrelación parcial a la vez

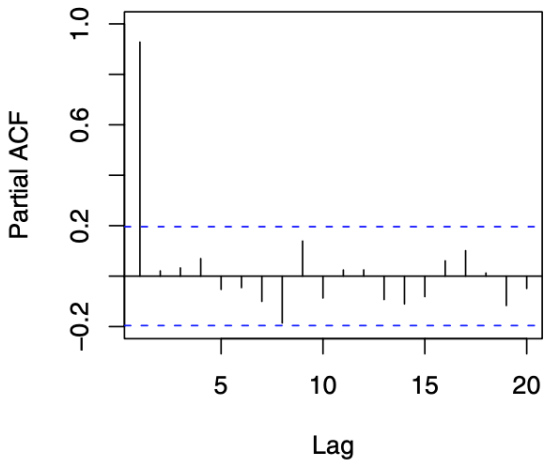
  - La función de autocorrelación parcial es similar a la de autocorrelación, pero esta no tiene en cuenta la correlación existente en los valores entre t y $t - s$, sino solo la del valor en t y en $t - s$

  - Para un modelo autorregresivo, la función de autocorrelación decrece de manera exponencial a cero, dado que las observaciones pasadas influyen en las observaciones presentes en el modelo, pero esta influencia disminuye cada vez más

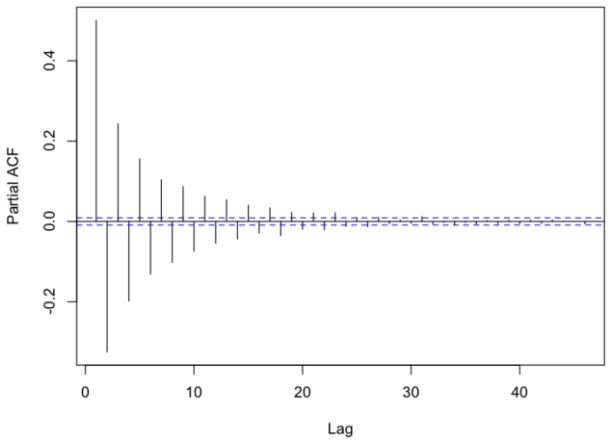


  - El número de retrasos significativos en la función de autocorrelación parcial (independientemente de la frecuencia de los datos) determinará el orden del proceso AR, ya que esos son

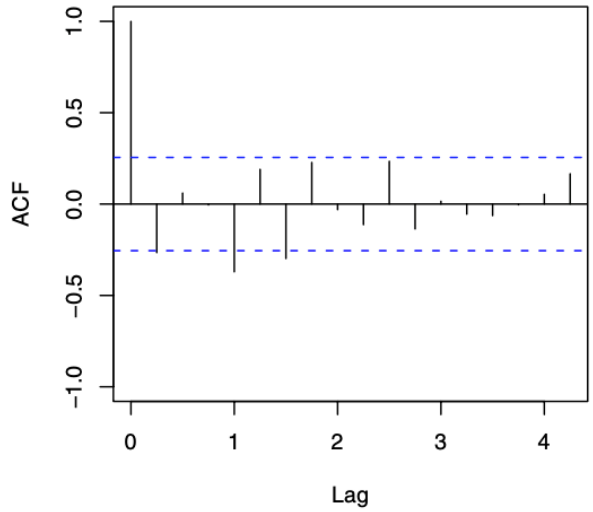
los retrasos que determinan significativamente más la autocorrelación



- Para un modelo de media móvil, la función de autocorrelación parcial decrece de manera exponencial a cero, dado que las observaciones pasadas no influyen en las observaciones presentes en el modelo solo se determina por los *shocks*, aunque estos influyen cada vez menos



- El número de retrasos significativos en la función de autocorrelación ((independientemente de la frecuencia de los datos) determinará el orden del proceso MA, ya que esos son los retrasos que determinan significativamente más la autocorrelación



- Con un proceso mixto como el $ARMA$ o el $ARIMA$, ambas funciones exhiben un descenso exponencial, por lo que no se pueden identificar un AR o un MA por separado. El orden, en este caso, vendrá determinado de otros modos (a través del uso de criterios de información u otros)

- Si a través del modelaje del término de error se consigue crear una serie temporal estacionaria, entonces los residuos del modelo estimado deberían comportarse como un proceso de ruido blanco

- En este caso, no se puede apreciar ningún patrón de tendencia o de estacionalidad en los residuos. Además, la mayoría de retrasos no son significativos cuando se mira la función de autocorrelación (solo algunos muy grandes, lo cual es común) y se distribuye como una normal (aunque haya algunas observaciones extremas debido a la muestra)

- Los modelos anteriores pueden ser caracterizados por su orden, el valor de sus parámetros y el comportamiento de sus autocorrelaciones, por lo que se pueden considerar varias formas en diferentes puntos. No obstante, el modelo $AR(1)$ es el modelo más utilizado con diferencia

    - Los procesos más complejos suelen ser muy difíciles de analizar, pero una razón más práctica para el uso de este modelo es que es muy optimista pensar que se puede saber precisamente la forma correcta para el modelo de las perturbaciones para cualquier situación

        - La autorregresión de primer orden ha demostrado a lo largo del tiempo que es un modelo razonable para el proceso subyacente de las perturbaciones, el cual en realidad seguramente es impenetrable

    - El modelo autorregresivo de primer orden es representado en su forma autorregresiva de la siguiente manera:

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

        - En este caso, se asume que el término de error en este proceso cumple las siguientes condiciones:

$$E(u_t|\boldsymbol{X}) = 0 \qquad Var(u_t|\boldsymbol{X}) = \sigma_u^2$$

$$Cov(u_t, u_s|\boldsymbol{X}) = 0 \quad for\ t \neq s$$

        - Debido a que u_t es un ruido blanco, los momentos condicionales son equivalentes a los momentos incondicionales, por lo que $E(u_t|\boldsymbol{X}) = E(u_t)$ y así

    - Por sustitución repetitiva, se puede ver que el proceso autorregresivo se puede expresar de la siguiente forma:

$$\varepsilon_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \cdots$$

        - A partir del proceso de media móvil, es evidente que cada perturbación ε_t carga con el historial entero de u, en donde las observaciones más recientes son las que tienen más peso y las más alejadas tienen menos. Los signos en este caso son positivos, aunque también se puede usar el signo opuesto sin cambiar las propiedades teóricas del modelo

        - Dado que estos valores de u_t no están correlacionados, la varianza de ε_t se puede expresar de la siguiente forma:

$$Var(\varepsilon_t) = \sigma_u^2 + \rho^2\sigma_u^2 + \rho^4\sigma_u^2 + \cdots$$

- Se impone la restricción de que $|\rho| < 1$ debido a que, de otra manera, la suma de la parte derecha de la ecuación anterior tiende a infinito (suposición de estacionariedad discutida anteriormente). Esto, a su vez, implica las siguientes propiedades para el proceso autorregresivo de primer orden:

$$\lim_{s\to\infty} \rho^s = 0 \qquad E(\varepsilon_t) = 0 \qquad Var(\varepsilon_t) = \frac{\sigma_u^2}{1-\rho^2}$$

- Con la suposición de estacionariedad, hay una forma fácil de obtener la varianza y la covarianza para t y $t-1$, debido a que $Cov(u_t, \varepsilon_s) = 0$ para $t > s$ y $Var(\varepsilon_t) = Var(\varepsilon_{t-1})$

$$Cov(u_t, \varepsilon_s) = Cov(u_t, u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \cdots) =$$

$$= Cov(u_t, u_s) + \rho Cov(u_t, u_{s-1}) + \rho^2 Cov(u_t, u_{s-2}) + \cdots = 0$$

$$\Rightarrow Var(\varepsilon_t) = \rho^2 Var(\varepsilon_{t-1}) + Var(u_t) + 2Cov(u_t, \varepsilon_s) =$$

$$= \rho^2 Var(\varepsilon_{t-1}) + \sigma_u^2 = \frac{\sigma_u^2}{1-\rho^2}$$

$$\Rightarrow Cov(\varepsilon_t, \varepsilon_{t-1}) = E(\varepsilon_t \varepsilon_{t-1}) = E[\varepsilon_{t-1}(\rho\varepsilon_{t-1} + u_t)]$$

$$= \rho Var(\varepsilon_{t-1}) = \frac{\rho\sigma_u^2}{1-\rho^2}$$

- Sustituyendo repetidamente, se puede obtener la siguiente expresión de ε_t en términos de las perturbaciones ε y de u para cualquier s:

$$\varepsilon_t = \rho^s \varepsilon_{t-s} + \sum_{i=0}^{s-1} \rho^i u_{t-i}$$

- A partir de esta, se puede obtener una expresión para la covarianza y para la correlación para cualquier t y $t-s$. Con la suposición de estacionariedad, se puede ver como la covarianza y la correlación se desvanece con el tiempo

$$Cov(\varepsilon_t, \varepsilon_{t-s}) = E[\varepsilon_t \varepsilon_{t-s}] =$$

$$= \rho^s Cov(\varepsilon_{t-s}, \varepsilon_{t-s}) + \sum_{i=0}^{s-1} \rho^i Cov(u_{t-i}, \varepsilon_{t-s}) = \rho^s Var(\varepsilon_{t-s}) = \frac{\rho^s \sigma_u^2}{1-\rho^2}$$

$$Corr(\varepsilon_t, \varepsilon_{t-s}) = \frac{Cov(\varepsilon_t, \varepsilon_{t-s})}{Var(\varepsilon_{t-s})} = \rho_s = \rho^s$$

- Dependiendo del signo de ρ, las autocorrelaciones pueden decrecer siguiendo una progresión geométrica o alternarse en signo si ρ es negativa

$$\sigma^2\boldsymbol{\Omega} = \frac{\sigma_u^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \rho \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & \rho & 1 \end{bmatrix}$$

Los datos no estacionarios

- La mayoría de variables económicas que exhiben tendencias son no estacionarias y, por tanto, no se puede utilizar el análisis anteriormente planteado porque presenta varias complicaciones

  - Los tres operadores más importantes a la hora de modelar series temporales es el operador de retraso L, el de primeras diferencias Δ, y el de diferencias estacionales Δ_s, los cuales simplifican las matemáticas del análisis. Estos se definen por las siguientes operaciones:

$$Ly_t \equiv y_{t-1}$$

$$\Delta y_t \equiv (1-L)y_t = y_t - y_{t-1} \qquad \Delta_s y_t \equiv (1-L^s)y_t = y_t - y_{t-s}$$

    - A partir de la definición del operador L, se puede ver como este cumple las siguientes propiedades:

$$L^p y_t = L\left(L\left(L(\ldots(Ly_t))\right)\right) = L\left(L\left(L(\ldots(y_{t-1}))\right)\right) = y_{t-p}$$

$$(L^p)^q y_t = L^{pq} y_t = y_{t-pq}$$

$$(L^p)(L^q)y_t = L^{p+q}y_t = L^p y_{t-q} = y_{t-q-p}$$

    - A partir de la definición del operador Δ, se puede ver como este cumple la siguiente propiedad:

$$\Delta^p = \Delta\left(\Delta\left(\ldots(\Delta(\Delta y_t))\right)\right) = \Delta\left(\Delta\left(\ldots(\Delta(y_t - y_{t-1}))\right)\right) =$$

$$= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) - \cdots - (y_{t-p+1} - y_{t-p})$$

- A partir de la definición del operador Δ_s, se puede ver como este cumple la siguiente propiedad:

$$\Delta_s Y_t = (1 + L + \cdots + L^{s-1})\Delta Y_t = y_t - y_{t-s}$$

- Finalmente, en una serie autorregresiva de primer orden con $|\beta| < 1$, se puede expresar y_t de la siguiente manera:

$$y_t = \beta y_{t-1} + \varepsilon_t \;\Rightarrow\; (1 - \beta L) y_t = \varepsilon_t$$

$$\Rightarrow\; y_t = \left(\frac{1}{1 - \beta L}\right)\varepsilon_t = (1 + \beta L + \beta^2 L^2 + \cdots)\varepsilon_t = \sum_{s=0}^{\infty} \beta^s \varepsilon_{t-s}$$

- Un proceso que figura comúnmente en situaciones de series temporales no estacionarias es el camino aleatorio o *random walk* con una tasa de deriva o *drift*, el cual tiene la siguiente forma:

$$y_t = \mu + y_{t-1} + \varepsilon_t$$

- Por sustitución directa, se puede ver como y_t es la suma infinita de perturbaciones y de la tasa de deriva (con una posible media diferente de cero). Si las innovaciones ε_t se generan por el mismo proceso con media nula y varianza constante, entonces la varianza de y_t sería infinita

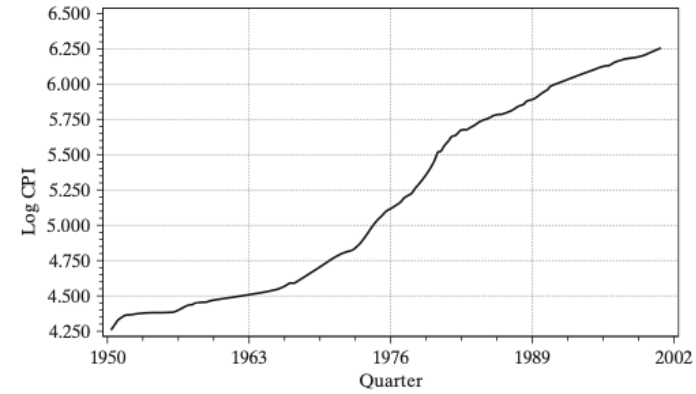
$$y_t = \frac{\mu + \varepsilon_t}{1 - L} = \sum_{s=0}^{\infty} (\mu + \varepsilon_{t-s})$$

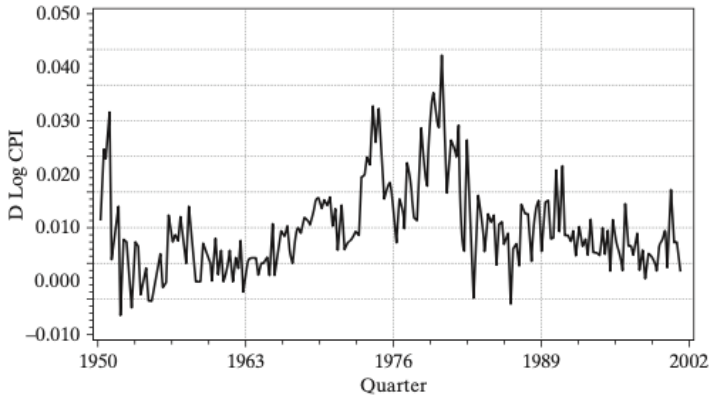
- Debido a que la varianza sería infinita, se puede ver como el proceso no sería un proceso estacionario, aunque $\mu = 0$. No obstante, el proceso de la primera diferencia Δy_t es el término de innovaciones más la media de la diferencia, de modo que es estacionario (media y varianza constante)

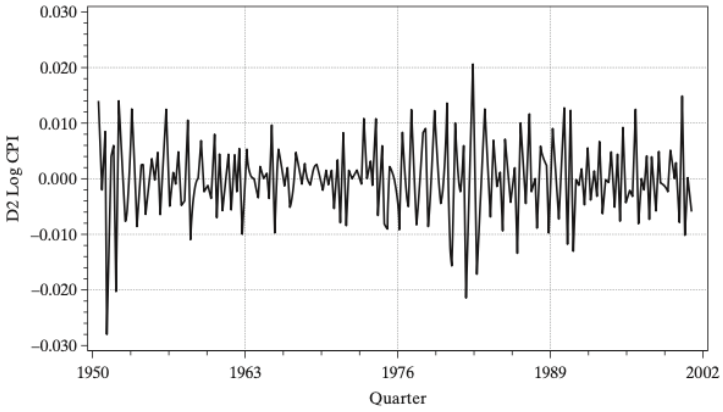
$$\Delta y_t = y_t - y_{t-1} = \mu + \varepsilon_t$$

$$E(\Delta y_t) = \mu \qquad Var(\Delta y_t) = \sigma_\varepsilon^2$$

- Por lo tanto, la serie temporal de y_t se denomina serie integrada de primer orden, denotado por $I(1)$, dado que tomar la primera diferencia produce un proceso estacionario. Una serie no estacionaria es una serie integrada de d-ésimo orden, denotado por $I(d)$, si tomar la primera diferencia d veces produce una serie estacionaria







- Una generalización del modelo $ARMA$ sería el modelo $ARIMA(p, d, q)$, el modelo autorregresivo integrado con media móvil, el cual se puede escribir de la siguiente forma:

$$\Delta^d y_t = (1-L)^d y_t = \mu + \gamma_1 \Delta^d y_{t-1} + \cdots + \gamma_p \Delta^d y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$$

  - Este resultado se puede escribir de manera compacta utilizando los polinomios $C(L)$ y $D(L)$ en el operador de retraso

$$C(L)[(1-L)^d y_t] = \mu + D(L)\varepsilon_t$$

$$C(L)x_t \equiv \left(1 - \gamma_1 L - \gamma_2 L^2 - \cdots - \gamma_p L^p\right)x_t$$

$$D(L)x_t \equiv \left(1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_q L^q\right)x_t$$

- En su investigación de 1974, Granger y Newbold discutieron sobre como la alta autocorrelación entre los residuos en los modelos de regresión convencionales pueden hacer que los contrastes de hipótesis fueran engañosos. Esto es causado por la relación entre la existencia de tendencia, la raíz unitaria y el camino aleatorio

    - Con este tipo de datos, los contrastes convencionales tenderán resultar en relaciones significativas entre variables cuando, en verdad, no hay ninguna relación

        - El resultado general es que la regresión lineal convencional de un proceso de camino aleatorio sobre otro sugeriría que existe una relación significativa, aunque estas dos sean realmente independientes. A este tipo de regresiones se las llama regresiones espurias

    - Para la mayoría de series temporales, el proceso de camino aleatorio con tasa de deriva y el proceso de tendencia estacionaria, en donde ε_t es ruido blanco en cada caso, son buenas caracterizaciones

$$y_t = \mu + y_{t-1} + \varepsilon_t \quad (random\ walk\ with\ drift)$$

$$y_t = \mu + \beta t + \varepsilon_t \quad (trend\ stationary)$$

        - Claramente, ambas regresiones producen series no estacionarias y con una tendencia fuerte, de modo que no es sorprendente que regresiones que involucren este tipo de variables casi siempre produzcan resultados significativos

        - Esta fuerte correlación parece ser debida a la tendencia subyacente, pero la intuición de este razonamiento es menos clara en el caso de un camino aleatorio puro, donde no hay una tasa de deriva que marque una tendencia clara (ni analíticamente ni gráficamente) pero las relaciones parecen persistir aún con series no relacionadas

$$y_t = y_{t-1} + \varepsilon_t \quad (pure\ random\ walk)$$

    - El factor común en estos tres tipos de procesos no es la tendencia, sino que el proceso generador de datos de los tres procesos se puede escribir como una ecuación con raíz unitaria:

$$(1 - L)y_t = \alpha + v_t$$

- Esta ecuación característica tiene una sola raíz equivalente a uno (cuando se considera L como variable), por lo que se le llama raíz unitaria

- El uso de datos caracterizados por raíces unitarias tiene el potencial de llevar a errores de inferencia serios, y estos datos pertenecen a series temporales no estacionarias

- Para los tres procesos, lo más natural sería comenzar diferenciando con tal de hacer a la serie estacionaria. No obstante, no es inmediatamente obvio como proceder correctamente, dado que no necesariamente se transforma la serie en estacionaria si se aplica un enfoque incorrecto

  - Hacer la primera diferencia para un proceso de camino aleatorio con tasa de deriva y/o puro permite obtener series estacionarias, pero hacer la primera diferencia para un proceso de tendencia estacionaria hace que se intercambie la tendencia por autocorrelación en forma de un proceso de media móvil de primer orden

$$\Delta y_t = \mu + \varepsilon_t \quad (random\ walk\ with\ drift)$$

$$\Delta y_t = \varepsilon_t \quad (pure\ random\ walk)$$

$$\Delta y_t = \beta + \varepsilon_t - \varepsilon_{t-1} \quad (trend\ stationary)$$

  - Por otro lado, extraer la tendencia (calcular los residuos de una regresión sobre el tiempo t) es contraproducente, mientras que para un proceso de tendencia estacionaria sería correcto

$$y_t - \dot{y}_t = \varepsilon_t - u_t \quad where \quad \dot{y}_t \equiv \alpha + \beta t + u_t$$

  - Dado que ninguno de estos enfoques es preferible de manera obvia, se puede considerar la primera diferencia en un modelo mixto que considere tanto una tendencia temporal como el primer retraso, y se puede ver como esta ecuación proporciona una base para tests de raíces unitarias con datos económicos

$$y_t = \mu + \beta t + \gamma y_{t-1} + \varepsilon_t$$

$$\Rightarrow \quad y_t - y_{t-1} = \mu + \beta t + (\gamma - 1) y_{t-1} + \varepsilon_t$$

  - En principio, un test de contraste de hipótesis para $\gamma - 1 = 0$ o, equivalentemente, para $\alpha_2 = 0$, da una confirmación de que el proceso es un camino aleatorio con tasa de deriva (si $\alpha_1 = 0$ también) y se puede aplicar la primera diferencia. Si $\gamma - 1 < 0$, entonces la evidencia favorece a un proceso de tendencia

estacionaria (o a otro modelo), y la extracción de la tendencia (o alguna alternativa) es preferible

- Las implicaciones de las raíces unitarias son profundas, dado que dependiendo de como es la serie temporal no estacionaria, el modelo permite inferir una u otra cosa

  - Si el modelo real de una variable es el de una serie integrada de primer orden (un camino aleatorio), entonces los *shocks* de las innovaciones son permanentes

  - Si el modelo real de una variable es el de una serie que sigue un proceso de tendencia estacionaria, los *shocks* de las innovaciones son temporales

- El único problema práctico con el tipo de test propuesto es que los procedimientos de inferencia estándar no son válidos cuando las series son estacionarias (las convergencias no se dan), de modo que se necesitan métodos alternativos

  - Considerando un modelo $AR(1)$ con media nula y con innovaciones que siguen un proceso de ruido blanco, se puede obtener que el sesgo del estimador de mínimos cuadrados cuando $\gamma \to 1$ es negativo

$$y_t = \gamma y_{t-1} + \varepsilon_t$$

$$\hat{\gamma}_{OLS} = \frac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2} \;\Rightarrow\; E(\hat{\gamma}_{OLS}) - \gamma < 0$$

    - Para $|\gamma| < 1$, en cambio, el estimador de mínimos cuadrados converge en probabilidad al verdadero parámetro γ y, por el teorema central del límite, se puede ver que la distribución del estadístico para γ sigue una distribución normal de media nula y varianza $1 - \gamma^2$

$$\hat{\gamma}_{OLS} = \frac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2} \;\Rightarrow\; \lim_{T\to\infty} P(|\hat{\gamma}_{OLS} - \gamma| > c) = 0$$

$$\Rightarrow\; \sqrt{T}(\hat{\gamma}_{OLS} - \gamma) \xrightarrow{D} N(0, 1 - \gamma^2)$$

  - Este resultado, es dudoso para el caso en el que $\gamma = 1$, ya que implicaría una distribución con varianza nula. No obstante, Dickey y Fuller, en su investigaciones de 1979 y 1981, mostraron la siguiente convergencia en distribución cuando $\gamma = 1$:

$$T(\hat{\gamma}_{OLS} - \gamma) \xrightarrow{D} v$$

- Esta convergencia expresa que el estadístico tiende a una variable v con varianza finita y positiva y, para muestras finitas, con un sesgo negativo, dado que $E(\hat{\gamma}_{OLS}) < 1$

- De este modo, los resultados obtenidos por Dickey y Fuller muestran como el estimador tiene un sesgo a la baja y como la convergencia de este estimador es más rápido a su valor poblacional, dado que la varianza bajo la hipótesis nula es $O(1/T^2)$ y no $O(1/T)$

- Las medidas convencionales tenderán a esconder el valor real de γ, dado que $\hat{\gamma}_{OLS}$ estará sesgado y, debido a la varianza tan pequeña que tendrá el estimador, los tests de hipótesis con el estadístico t tenderán a rechazar la hipótesis nula de que $\gamma = 1$

- Por lo tanto, ambos propusieron seguir usando este estadístico, pero revisar los valores para los cuales se rechaza la hipótesis de que $\gamma = 1$, los cuales ambos propusieron al obtenerlos a través de simulación de Monte Carlo

- Para poder contrastar la hipótesis de raíces unitarias, se pueden utilizar diferentes versiones del test de Dickey-Fuller dependiendo del tipo de serie temporal

  - En todas las versiones siguientes se asume que el término de error se distribuye normalmente y que no hay covarianza intertemporal entre dos realizaciones del término

$$\varepsilon \sim N(0, \sigma^2) \qquad Cov(\varepsilon_t, \varepsilon_{ss}) = 0 \ \ for \ \ \forall t \neq s$$

  - La versión más simple del modelo a analizar es la de un camino aleatorio

$$y_t = \gamma y_{t-1} + \varepsilon_t$$

  - No obstante, un camino aleatorio simple suele ser inadecuado para muchas series, de modo que se planea utilizar un camino aleatorio con tasa de deriva

$$y_t = \mu + z_t \ \ where \ \ z_t = \gamma z_{t-1} + \varepsilon_t$$

$$\Rightarrow \ y_t = \mu(1-\gamma) + \gamma z_{t-1} + \varepsilon_t$$

  - No obstante, un camino aleatorio simple suele ser inadecuado para muchas series, de modo que se planea utilizar un camino aleatorio con tasa de deriva

$$y_t = \mu + \beta t + z_t \quad where \quad z_t = \gamma z_{t-1} + \varepsilon_t$$

$$\Rightarrow y_t = [\mu(1-\gamma) + \gamma\beta] + \beta(1-\gamma)t + \gamma y_{t-1} + \varepsilon_t$$

- Bajo la hipótesis nula de que $\gamma = 1$, se puede utilizar el estadístico t para γ para cualquiera de las versiones vistas, solo que se tiene que tener en cuenta los valores críticos correspondientes para cada caso

$$DF = \frac{\hat{\gamma} - 1}{Est.Std.Error(\hat{\gamma})}$$

	Sample Size			
	25	**50**	**100**	**∞**
F ratio (D–F)[a]	7.24	6.73	6.49	6.25
F ratio (standard)	3.42	3.20	3.10	3.00
AR model[b] (random walk)				
0.01	−2.66	−2.62	−2.60	−2.58
0.025	−2.26	−2.25	−2.24	−2.23
0.05	−1.95	−1.95	−1.95	−1.95
0.10	−1.60	−1.61	−1.61	−1.62
0.975	1.70	1.66	1.64	1.62
AR model with constant (random walk with drift)				
0.01	−3.75	−3.59	−3.50	−3.42
0.025	−3.33	−3.23	−3.17	−3.12
0.05	−2.99	−2.93	−2.90	−2.86
0.10	−2.64	−2.60	−2.58	−2.57
0.975	0.34	0.29	0.26	0.23
AR model with constant and time trend (trend stationary)				
0.01	−4.38	−4.15	−4.04	−3.96
0.025	−3.95	−3.80	−3.69	−3.66
0.05	−3.60	−3.50	−3.45	−3.41
0.10	−3.24	−3.18	−3.15	−3.13
0.975	−0.50	−0.58	−0.62	−0.66

- Los contrastes de Dickey-Fuller asumen que las perturbaciones del modelo son ruido blanco, por lo que, si no se cumple la suposición, este contraste no es adecuado. No obstante, se puede extender al contraste aumentado de Dickey-Fuller, el cual se basa en el siguiente modelo:

$$y_t = \mu + \beta t + \gamma y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_p \Delta y_{t-p} + \varepsilon_t$$

  - El estadístico de contraste para este modelo es el mismo que antes, de modo que solo se tienen que considerar los valores críticos correctos

$$DF = \frac{\hat{\gamma} - 1}{Est.Std.Error(\hat{\gamma})}$$

  - El camino aleatorio simple se obtiene imponiendo $\mu = 0$ y $\beta = 0$, mientras que si $\beta = 0$, el camino aleatorio tiene una tasa de deriva

- Una alternativa sugerida para mejorar las propiedades del contraste para muestras finitas o de generalización es el contraste de Phillips-Perron (PP). Este se basa en el siguiente modelo, en donde δ_t puede ser 0, μ o $\mu + \beta t$:

$$y_t = \delta_t + \gamma y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_p \Delta y_{t-p} + \varepsilon_t$$

  - El procedimiento modifica los estadísticos de Dickey-Fuller para obtener el siguiente estadístico, llamado estadístico PP:

$$Z = \sqrt{\frac{c_0}{a}}\left(\frac{\hat{\gamma} - 1}{v}\right) - \frac{1}{2}(a - c_0)\frac{Tv}{\sqrt{as^2}}$$

$$where$$

$$s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T - K} \quad \& \quad v^2 = estim.asympt.variance\ of\ \hat{\gamma}$$

$$c_j = \frac{1}{T}\sum_{s=j+1}^{T} e_t e_{t-s} \quad for\ j = 0, \ldots, L \quad \& \quad c_0 = \left(\frac{T - K}{T}\right)s^2$$

$$a = c_0 + 2\sum_{j=1}^{L}\left(1 - \frac{j}{L+1}\right)c_j$$

  - Como se puede ver, este tipo de contraste utiliza las mismas ecuaciones que el ADF pero propone una corrección no paramétrica del estadístico para permitir la autocorrelación de los residuos

  - Igual que el ADF, tampoco sigue una distribución t-Student, por lo que los valores críticos también se tienen que derivar a partir de simulaciones de Monte Carlo. No obstante, como las ecuaciones y las hipótesis nulas son las mismas, la distribución es la misma y se utilizan los mismos valores críticos que con el ADF

- Otra alternativa también utilizada es el contraste KPSS, el cual permite contrastar la hipótesis nula de estacionariedad en un modelo con ε_t estacionaria y Z_t es una serie estacionaria con media nula y varianza uno en el siguiente modelo:

$$y_t = \alpha + \beta t + \gamma Z_t + \varepsilon_t \quad for\ t = 1,2, \ldots, T$$

  - Si $\gamma = 0$, entonces el proceso es estacionario si $\beta = 0$ y es estacionario con tendencia si $\beta \neq 0$. Como Z_t es $I(1)$, entonces y_t es no estacionaria si γ no es cero

- El test KPSS para la hipótesis nula de que $\gamma = 0$ contra la alternativa de que no lo es invierte la estrategia del estadístico de Dickey-Fuller, lo cuál permite que se tenga una mayor potencia en el contraste de hipótesis cuando γ está muy cerca de 0 dado que se usa un estadístico estacionario y no uno no estacionario como el de ADF o PP

- Bajo la hipótesis nula, α y β solo se pueden estimar por MCO

- El estadístico de contraste KPSS es el siguiente, donde L se escoge a conveniencia:

$$KPSS = \frac{\sum_{t=1}^{T} E_t^2}{T^2 \hat{\sigma}^2} \quad where \quad E_t = \sum_{s=1}^{t} e_s$$

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{T} e_t^2}{T} + 2\sum_{j=1}^{L}\left(1 - \frac{j}{L+1}\right) r_j \qquad r_j = \frac{\sum_{s=j+1}^{T} e_s e_{s-j}}{T}$$

- Los estudios económicos con series temporales casi siempre involucran series no estacionarias y tendenciosas, de modo que se puede utilizar la diferencia o transformaciones de otros tipos para poder obtener estacionariedad. Sin embargo, existen otros métodos más apropiados para analizar este tipo de variables

    - En un modelo completamente especificado $y_t = \beta x_t + \varepsilon_t$, se suele asumir que el proceso de las perturbaciones ε_t es un ruido blanco estacionario

        - No obstante, de manera general, si dos series están integradas en diferentes órdenes, entonces las combinaciones lineales estarán integradas al mayor de ambos órdenes

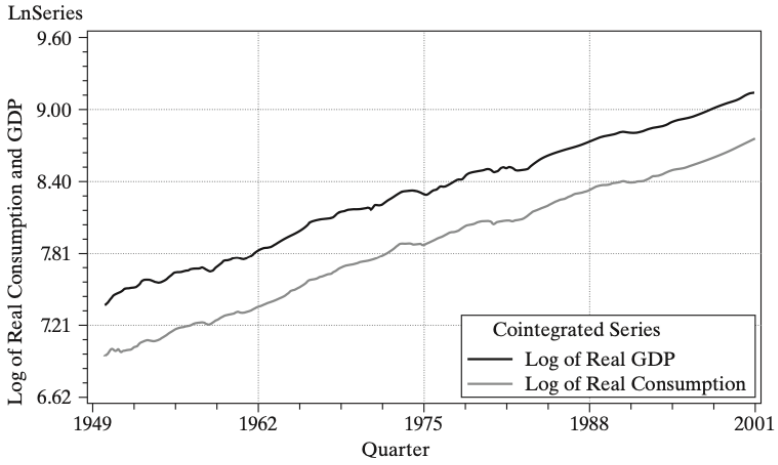
        - Por lo tanto, si y_t y x_t son series integradas $I(1)$, entonces normalmente se esperaría que $y_t - \beta x_t$ también fuera $I(1)$ independientemente del valor de β, y no $I(0)$

        - Si y_t y x_t tienen una tasa de deriva (en la misma dirección) para sus propias tendencias, a no ser que haya una relación entre sus tendencias, la diferencia entre estas debe ser también creciente con otra tendencia, y esto causaría que la suposición sobre ε_t fuera inconsistente

    - No obstante, también puede darse el caso de que si y_t y x_t son series integradas $I(1)$, de modo que puede existir un valor β tal que $\varepsilon_t = y_t -$

βx_t sea $I(0)$ (existe una diferencia concreta $y_t - \beta x_t$ que cree una serie estacionaria)

- Intuitivamente, si las dos series son $I(1)$, entonces esta diferencia parcial entre ellas puede ser estable alrededor de una media fija. La implicación sería que las series van acercándose a la misma velocidad (más o menos)



- Dos series que satisfacen estos requerimientos se dicen que están cointegradas (su combinación lineal está cointegrada), y el vector $(1, -\beta)$ (o cualquier múltiplo de este) es un vector de cointegración

- En este caso, se distingue entre una relación a largo plazo entre y_t y x_t (la manera en que dos variables se desplazan juntas hacia arriba) y las dinámicas a corto plazo (la relación entre desviaciones de y_t de su tendencia a largo plazo y de x_t de su tendencia a largo plazo)

- De estar cointegradas, la diferenciación de los datos sería contraproductiva, dado que oscurecería la relación a largo plazo entre y_t y x_t

- Si dos variables $I(1)$ están cointegradas, entonces alguna combinación lineal ($y_t - \beta x_t$) de estas variables produce una serie estacionaria $I(0)$ (la combinación lineal está cointegrada). Esta combinación lineal no crea misteriosamente una variable estacionaria, sino que hay algo presente en las variables originales que ya no debe de estar en la nueva

  - Suponiendo que hay dos variables $I(1)$ que tienen una tendencia lineal (donde u_t y v_t es ruido blanco), una combinación lineal de y_{1t} y y_{2t} con el vector $(1, \theta)$ produce la siguiente variable z_t que es, generalmente, $I(1)$:

$$y_{1t} = \alpha + \beta t + u_t \qquad y_{2t} = \gamma + \delta t + v_t$$

$$\Rightarrow z_t = (\alpha + \theta\gamma) + (\beta + \theta\delta)t + u_t + \theta v_t$$

- La única manera de que z_t pueda hacerse estacionaria es que $\theta = -\beta/\delta$. De ser así, entonces el efecto de una combinación lineal de ambas variables es que se elimine la tendencia lineal común entre ellas

- No obstante, esto implica que la única manera de que y_{1t} y y_{2t} estén cointegradas es que tengan una tendencia común de algún tipo. Si, además de la tendencia lineal se incluyera una tendencia estocástica a través de un camino aleatorio $w_{it} = w_{it-1} + \eta_{it}$ en donde η_t para $i = 1{,}2$ es ruido blanco, entonces las series solo estarían cointegradas si $w_{1t} = w_{2t}$ (dado que una combinación lineal no implicaría caminos aleatorios diferentes, solo uno)

$$y_{1t} = \alpha + \beta t + \lambda w_t + u_t \qquad y_{2t} = \gamma + \delta t + \pi w_t + v_t$$

- Se puede demostrar, sin embargo, que no es posible encontrar una combinación lineal que esté cointegrada aunque compartan la misma tendencia estocástica. Por lo tanto, el resultado final es que si y_{1t} y y_{2t} están cointegradas, entonces tienen que compartir una sola tendencia común (no dos comunes, como en este caso)

- Stock y Watson determinaron que esta última conclusión era la más importante en el caso de las variables económicas. Un conjunto de M variables que están cointegradas se puede escribir como un componente estacionario sumado a combinaciones lineales de un conjunto más pequeño de tendencias comunes. Si el rango de cointegración de un sistema es r, solo puede haber como mucho $M - r$ tendencias lineales y $M - r$ tendencias estocásticas comunes. El efecto de la cointegración es purgar estas tendencias comunes de las variables resultantes

- Para analizar la cointegración, el primer paso es establecer si de verdad esta característica está presente en los datos. Hay dos enfoques para ello: el enfoque de Engle-Granger y el del VAR

  - El enfoque de Engle-Granger consiste en contrastar la estacionariedad de los errores de equilibrio de estimaciones de una sola ecuación. El contraste que se utiliza en este enfoque es el contraste aumentado de Dickey-Fuller de Engle-Granger

    - Siendo $\boldsymbol{y_t}$ un conjunto de M variables que se cree que están cointegradas, el primer paso es contrastar el orden de

integración de cada una a través de un contraste de Dickey-Fuller estándar o de uno aumentado

- Si el rango de cointegración de un sistema es r, entonces hay r vectores independientes $\boldsymbol{\gamma_i} = (1, -\boldsymbol{\theta_i})$, en donde cada vector se distingue por ser normalizado en una variables diferente. Si se supone que hay un conjunto de variables exógenas $I(0)$, incluyendo un término constante en el modelo, entonces cada vector cointegrante produce la siguiente relación de equilibrio:

$$\boldsymbol{y_t'\gamma_i} = \boldsymbol{x_t'\beta} + \varepsilon_{it} \;\Rightarrow\; y_{it} = \boldsymbol{Y_{it}'\theta_i} + \boldsymbol{x_t'\beta} + \varepsilon_{it}$$

- Se pueden obtener estimadores por mínimos cuadrados ordinarios de esta regresión, y si la teoría es correcta y los estimadores son consistentes, entonces los residuos de esta regresión deberían estimar los errores de equilibrio

- Bajo estas suposiciones, los residuos de la regresión son estimadores de los errores de equilibrio y, por tanto, tendrían que ser $I(0)$. Aunque el enfoque natural sería aplicar los contrastes de Dickey-Fuller a estos residuos, se tienen que utilizar otros valores especiales, dado que la distribución no es normal y depende del número de variables $I(1)$ usadas como regresores