Apunte econometría I

Christian González I.

March 26, 2022

Abstract

Este apunte tiene como fin resumir lo leído y lo visto en clase de una manera más compacta, no obstante por ningún motivo trata de reemplazar ni asimilar la calidad que tiene el leer un libro econométrico o asistir a una clase.

1 ¿Qué es la econometría?

El término 'econometría' se piensa que fue creado por Ragnar Frish (1895-1973), el cual argumentaba de que la econometría, en resumidas cuentas, busca la *unificación* de la teoría económica y los datos cuantitativos. Esto es relevante al momento de hablar de correlación y causalidad, debido a que ambos conceptos pueden ser muy distintos, desde un punto de vista teórico/racional. Esto debido a que puede ser que dos variables presenten una alta correlación, pero no necesariamente implica que exista causalidad entre una y otra. A esto se le conoce como *correlación espuria*, para profundizar en este concepto supongamos de que tenemos dos observaciones:

$$\{y_t, x_t\}_{t=1}^{T} \tag{1}$$

Si expresamos los datos con desviaciones respecto a la media, tendremos lo siguiente:

$$\tilde{z_t} = z - \bar{z}, \quad \bar{z} = T^{-1}\sum_{t=1}^{T} z_t, \quad z = y, x \tag{2}$$

$$cov(y, x) = \mathbb{E}(yx) - \mathbb{E}(y)\mathbb{E}(x) \tag{3}$$

Como se puede observar, el signo de la covarianza lo determinará la correlación, además tendremos de que la covarianza depende de las unidades con que estén medidas las variables. Por otro lado, tendremos los denominados **coeficientes de correlación** los cuales no dependen de la unidad con que estén medidas las variables:

$$r = T^{-1}\sum_{t=1}^{T} \frac{\tilde{x}_t}{S_x}\frac{\tilde{y}_t}{S_y} = \frac{S_{xy}}{S_x S_y} \quad \text{con} \quad S_z = \sqrt{\sum_{t=1}^{T} \tilde{z}_t^2}, \quad z = y, x \tag{4}$$

Por Cauchy-Schwarz sabemos de que $-1 \leq r \leq 1$, cuando toma valor 1 es correlación perfecta, que es la mayoría de veces que la relación es determinística.
Como se puede observar en la ecuación 4 al dividir las variables por S_z se estandariza la unidad de medida, no obstante los coeficientes de correlación solo miden la asociación y cercanía lineal de las variables, no dicen nada de la relación causal[1], es importante lo explicado por anterioridad, porque puede darse el caso en que dos variables se relacionen de una manera no lineal.

[1] A la falacia de creer lo contrario se le conoce como ¨*Post hoc, ergo propter hoc*¨ y no es poco común ver en la sociedad.

1.1 El marco de regresión

A continuación hablaremos el marco de regresión, para lo cual asumiremos de que tenemos el siguiente vector de variables $w_t = \{w_1, w_2, \ldots, w_t\}$, supongamos de que tenemos 2 particiones (ambas variables aleatorias) del estilo:

$$w_t = (y_t, x_t) \quad \forall \quad y_t \in \mathbb{R}, \quad x_t \in \mathbb{R}^k \tag{5}$$

A continuación denotaremos la función de probabilidad conjunta (para lo cual asumiremos de que ambas variables son continuas), la cual muestra la probabilidad de encontrar ambos eventos:

$$f(y_t, x_t, \theta) \tag{6}$$

Donde θ es un parámetro desconocido, no obstante por teorema de bayes sabemos de que la función conjunta es igual al producto entre la marginal y la condicional, llegando a lo siguiente:

$$f(y_t, x_t, \theta) = f(y_t, \theta_1) \cdot f(x_t, \theta_2) \tag{7}$$

Cuando se puede hacer inferencia válida en θ_2 sin preocuparse de la marginal, es cuando θ_1 y θ_2 son libres de variación y por ende la marginal ($f(x_t, \theta_2)$ que es la que nos interesa) es libre de variación.
¿Cómo saber cual es el orden con que van las variables? esta respuesta dependerá del escenario en que nos encontremos, no obstante típicamente la teoría económica nos explica la relación entre las variables y la causalidad de los fenómenos, definimos:

- y : Es la variable dependiente o endógena.
- x : Es la variable independiente o exógena.

En estadística (y en econometría) nos basta con tener los primeros dos momentos de cada serie (media, varianza), ya que los momentos de mayor orden están compuestos en base a estos dos, teniendo esto presente e ignorando la marginal, dado de que asumiremos de que es libre de variación, tendremos que la media condicional y la varianza condicional son respectivamente:

$$m(x_t, \theta_3) = \mathbb{E}(y_t | x_t, \theta_3) = \int_{-\infty}^{\infty} y f(y | x_t, \theta_2) dy \tag{8}$$

$$g(x_t, \theta_4) = \int_{-\infty}^{\infty} y^2 f(y | x_t, \theta_2) dy - [m(x_t, \theta_3)]^2 \tag{9}$$

Es importante decir de que $m(x_t, \theta_3)$ puede ser no lineal. Teniendo la ecuación 8, podemos denotar la relación de causalidad de la forma:

$$y_t = m(x_t, \theta_3) + u_t \longrightarrow u_t = y_t - m(x_t, \theta_3) \tag{10}$$

Donde u_t es idependiente de todo y representa la diferencia entre la variable dependiente y la media **poblacional**. El error poblacional tiene las siguientes características[2]:

1. $\mathbb{E}(u_t, x_t) = 0$, demostración:

$$\mathbb{E}(u_t | x_t) = \mathbb{E}(y_t - m(x_t) | x_t) = \mathbb{E}(y_t | x) - \mathbb{E}(m(x_t) | x_t) = m(x_t) - m(x_t) = 0$$

2. Para la segunda propiedad, tenemos que introducir el concepto de esperanza iterada, la cual dice que si $\mathbb{E}|y| < \infty$ entonces para cualquier vector aleatorio de x, se cumple que:

$$\mathbb{E}(\mathbb{E}(y|x)) = \mathbb{E}(y) \tag{11}$$

   A esta se le conoce como la Ley simple de las expectativas iteradas, la cual nos dice, en resumidas cuentas, que el promedio de los promedios condicionales es el promedio incondicional.
   Cuando x es discreta:

$$\mathbb{E}(\mathbb{E}(y|x)) = \sum_{j=1}^{\infty} \mathbb{E}(y|x_j) Pr(x = x_j)$$

[2]Estas son por definición, no por supuestos.

Cuando x es continua:

$$\mathbb{E}(\mathbb{E}(y|x)) = \int_{\mathbb{R}^k} \mathbb{E}(y|x) f_x(x) dx$$

La Ley general de las esperanzas iteradas permite dos condiciones para las variables establecidas, la Ley de la experanza iterada nos dice que:
Si $\mathbb{E}|y| < \infty$ entonces para cualquier vectores x_1 y x_2 aleatorios, se cumple de que:

$$\mathbb{E}(\mathbb{E}(y|x_1, x_2)|x_1) = \mathbb{E}(y|x_1)$$

Esta nos dice, en resumidas cuentas que la variable que contenga la mayor cantidad de información es la que termina ganando.
Teniendo esto presente, se cumple la segunda propiedad la cual dice que $\mathbb{E}(u_t) = 0$, demostración:

$$\mathbb{E}(u_t) = \mathbb{E}(\mathbb{E}(u_t|x_t)) = \mathbb{E}(0) = 0$$

3. La tercera propiedad, nos dice que sea $h(x_t)$ una función cualquiera de x tal que $\mathbb{E}|h(x_t)u_t| < \infty$, la cual puede ser lineal o no lineal, por lo tanto se cumple de que $\mathbb{E}(h(x_t)u_t) = 0$, demostración:

$$\mathbb{E}(h(x_t)u_t) = \mathbb{E}(\mathbb{E}(h(x_t)u_t|x_t)) = \mathbb{E}(h(x_t)\mathbb{E}(u_t|x_t)) = \mathbb{E}(h(x_t) \cdot 0) = 0$$

4. Se cumple de que $\mathbb{E}(x_t u_t)$ son ortogonales, dado de que la esperanza es 0.

5. Si $\mathbb{E}|y_t|^r < \infty$ para $r > 1$, entonces $\mathbb{E}|u_t|^r < \infty$.

Es importante señalar que todo lo anterior se cumple por estructura (como está definido el error, que es la diferencia entre el verdadero parámetro Y y la esperanza de este) y no por modelo, es decir se hace realidad por definición, además de los primeros dos momentos $m(\cdot)$ y $g(\cdot)$ pueden tomar cualquier forma, cuando $m(\cdot)$ es lineal se dice que es un modelo de regresión lineal (LRM por sus siglas en inglés).

2 El modelo de regresión lineal (LRM)

En la sección anterior, explicamos lo que era el marco de regresión, no obstante nosotros típicamente nunca podremos encontrar un marco que asimile el comportamiento de los datos y la relación de causalidad al 100 %, esto debido a que hay limitaciones en la cantidad (y calidad) de los datos con que uno trabaja. No obstante, la econometría nos da estructuras con las que nosotros podemos crear modelos, los cuales son aproximaciones a la realidad. A continuación, mostraremos un modelo simple llamado modelo de regresión lineal (LRM) para lo cual asumiremos de que $m(x_t, \beta) = x_t' \cdot \beta$, es decir que $m(x_t, \beta)$ es una aproximación lineal, por lo tanto nuestro modelo se puede escribir de manera vectorial como sigue a continuación:

$$y_t = x_t' \cdot \beta + u_t \tag{12}$$

No obstante, si tenemos varias observaciones lo podemos denotar de la manera:

$$Y = X\beta + u \tag{13}$$

En donde:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}_{T\times 1}, \quad X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T,1} & x_{T,2} & \cdots & x_{T,k} \end{bmatrix}_{T\times k} = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_T' \end{bmatrix}_{T\times K}, \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}_{T\times 1}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}_{k\times 1} \tag{14}$$

En donde hay T sujetos (pueden ser personas, países, etc) que tiene k características, es importante destacar de que el error u, no es igual al del marco de regresión, debido a que puede estar correlacionado con los x (este problema se verá más adelante), lo que mostraremos a continuación son los supuestos que le haremos al modelo de regresión lineal, los cuales serán levantados en incisos más adelante, estos son:

1. $\mathbb{E}(u_t|x_t) = 0$, en este caso es supuesto y por ende es testeable.

2. $rank(X) = K$, es decir los X presentan rango completo (y por lo tanto, no presenta multicolinealidad perfecta, tema que veremos más adelante) y por ende $\det(X'X) \neq 0$, es invertible.

3. $\mathbb{E}(u_t u_s) = 0 \quad \forall t \neq s$, a este supuesto se le conoce como *ruido blanco* y nos dice que los errores pasados, no nos sirve para explicar el presente y por ende, no nos explica nada del futuro.
   Si añadimos otros supuestos adicionales (poco realistas), llegamos a lo que se conoce como el modelo de regresión lineal homocedástico (HLRM), el cual nos dice de que:

4. $\mathbb{E}(u_t^2|x_t) = \sigma^2$ o $\mathbb{E}(uu'|X) = \sigma^2 I_t$, es decir los errores son homocedástico o de igual varianza, gráficamente:

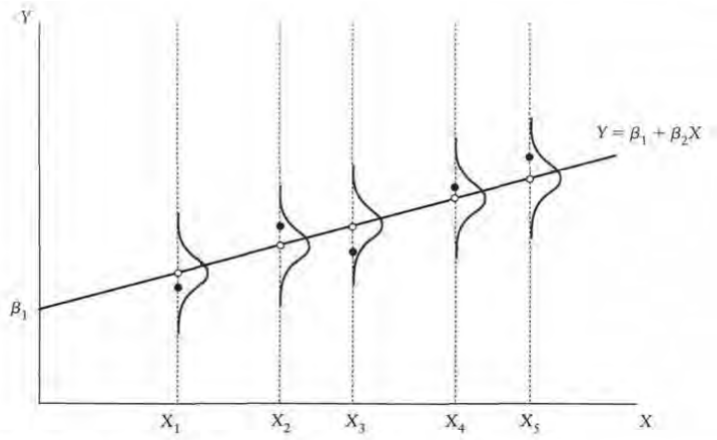


Figure 7.1 Homoscedasticity

   Finalmente, veremos el modelo de regresión lineal normal, el cual incluye un supuesto adicional sobre la distribución de los errores:

$$u_t \sim N(0, \sigma^2) \tag{15}$$

   Es importante señalar, que la normalidad del modelo la dan los residuos.

Estos supuesto, son poco realistas y casi no se presentan en los datos, no obstante hay soluciones que veremos más adelante a la hora de levantar estos supuesto.

2.1 Notas importantes : El mejor predictor

Como vimos en el primer inciso, nosotros teniendo una serie de valores realizados de x_t deseamos crear una predicción de y_t. Si escribimos cualquier predictor como una función $g(x_t)$ de x_t. Por estructura, el error de predicción será denotado como la diferencia entre $y_t - g(x_t)$. Una medida no estocástica de la magnitud de los errores de predicción es la esperanza de estos al cuadrado:

$$\mathbb{E}(y_t - g(x_t))^2 \tag{16}$$

Teniendo de que $y_t = m(x_t) + u_t$:

$$\mathbb{E}(m(x_t) + u_t - g(x_t))^2 = \mathbb{E}u_t^2 + 2\mathbb{E}(u_t(m(x_t) - g(x_t)) + \mathbb{E}(m(x_t) - g(x_t))^2 \tag{17}$$

Por definición de los u_t, tendremos de que $\mathbb{E}(\underbrace{u_t(m(x_t) - g(x_t))}_{=\mathbb{E}(u_t h(x_t))}) = 0$, por lo tanto:

$$\mathbb{E}(m(x_t) + u_t - g(x_t))^2 = \mathbb{E}u_t^2 + \mathbb{E}(m(x_t) - g(x_t))^2 \geq \mathbb{E}u_t^2 \tag{18}$$

$$\therefore \quad \mathbb{E}(y_t - g(x_t))^2 \geq \mathbb{E}(y_t - m(x_t))^2 \tag{19}$$

En donde: $m(x_t) = \mathbb{E}(y_t|x_t)$, esto demuestra que el mejor predictor es la media incondicional $\mu = \mathbb{E}(y)$. No obstante, si bien es el mejor predictor, típicamente no conocemos su forma funcional debido a que no conocemos su parámetros θ, por lo tanto introduciremos a lo que es denominado el mejor predictor lineal (tiene apellido, es algo más acotado) de los y_t dados los x_t, en donde asumiremos de que:

1. $\mathbb{E}y_t^2 < \infty$.
2. $\|x_t\|^2 < \infty$.
3. $\mathbb{E}(e_t^2) < \infty$
4. $Q_{xx} = \mathbb{E}(x_t x_t')$ es definida positiva.
5. Los momentos $\mathbb{E}(x_t x_t')$ y de $\mathbb{E}(x_t y_t')$ existen con elementos finitos.

Por lo tanto, el mejor estimador lineal será de la forma:

$$\mathcal{P}(y_t|x_t) = x_t'\boldsymbol{\beta} \tag{20}$$

Donde $\boldsymbol{\beta}$ minimiza la media al cuadrado de la predicción de los cuadrados, teniendo:

$$S(\boldsymbol{\beta}) = \mathbb{E}(y_t - x_t'\boldsymbol{\beta})^2 = \mathbb{E}y_t^2 - 2\boldsymbol{\beta}'\mathbb{E}(x_t y_t) + \boldsymbol{\beta}'\mathbb{E}(x_t' x_t)\boldsymbol{\beta} \tag{21}$$

Por lo tanto:

$$\boldsymbol{\beta} = \underset{\boldsymbol{\beta}\in\mathbb{R}^k}{\arg\min}\, S(\boldsymbol{\beta}) \tag{22}$$

La CPO es:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \longrightarrow -2\mathbb{E}(x_t y_t) + 2\mathbb{E}(x_t x_t')\boldsymbol{\beta} = 0 \tag{23}$$

Por lo tanto:

$$\boldsymbol{\beta} = \left[\mathbb{E}(x_t x_t')\right]^{-1} \mathbb{E}(x_t y_t) \tag{24}$$

Con esto tenemos una manera explícita de la proyección lineal de los y en los x:

$$\mathcal{P}(y_t|x_t) = x_t'\boldsymbol{\beta} = x_t' \left[\mathbb{E}(x_t' x_t)\right]^{-1} \mathbb{E}(x_t y_t) \tag{25}$$

Tendremos la siguiente propiedad:

$$\mathbb{E}(x_t e_t) = \mathbb{E}(x_t(y_t - x_t'\boldsymbol{\beta}) = \mathbb{E}(x_t y_t) - \mathbb{E}(x_t x_t'\boldsymbol{\beta}) = \mathbb{E}(x_t y_t) - \mathbb{E}(x_t x_t')\boldsymbol{\beta} \tag{26}$$

$$= \mathbb{E}(x_t y_t) - \underbrace{\mathbb{E}(x_t x_t') \left[\mathbb{E}(x_t x_t')\right]^{-1}}_{=I} \mathbb{E}(x_t y_t) = 0 \tag{27}$$

En el caso, de que el vector x_t tenga constante, va a implicar de que:

$$\mathbb{E}(e_t) = 0 \tag{28}$$

En el caso de que el vector regresor no contenga constante, esto no se garantiza. Esta es una razón para que la regresión siempre contenga constante, ya que de esta manera se asegura de que la media de la proyección del error sea igual a 0.
Cabe señalar de que antes teniamos:

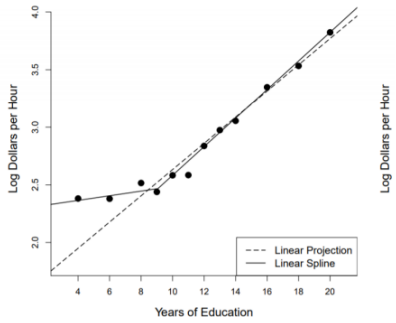
$$u_t = y_t - \underbrace{\mathbb{E}(y_t|x_t)}_{\text{Esperanza condicional}} \tag{29}$$

Que representaba a mi mejor predicción. No obstante, dabo este nuevo escenario tenemos:

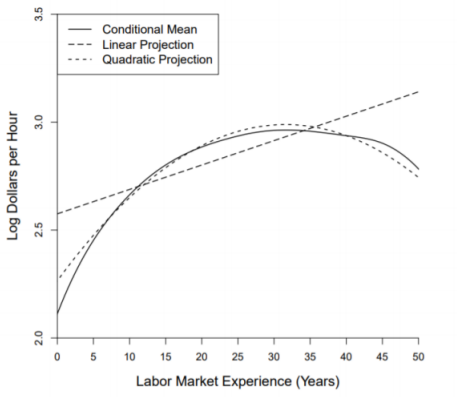
$$e_t = y_t - \underbrace{x_t'\boldsymbol{\beta}}_{\text{Predicción lineal}} \tag{30}$$

No obstante, el modelo lineal tiene una serie de problemas:

- No ajusta bien los datos si están dispersos, por ejemplo si vemos el logaritmo de los salarios contra los niveles de educación tendremos lo siguiente:



Como se puede observar, la linea punteada es la aproximación lineal la cual no se ajusta correctamente a los datos que se encuentran fuera de dicha recta. Una manera de solucionar este problema, es incluir dummies por rango de años de educación esta recta se llama spline lineal. Esto es relativamente efectivo, ya que con vectores lineales podemos hacer una buena aproximación a la media condicional:



- Sabemos de que $\mathbb{E}(x_t u_t) = 0$ (donde $u_t = y_t - \mathbb{E}(y_t|x_t)$). De modo que si la esperanza condicional (que es lo que buscamos) es lineal, el estimador va a ser el mejor proyector lineal. Sin embargo lo contrario no va a ser cierto, ya que el error de proyección $e_t = y_t - x_t'\beta$, no necesariamente va a cumplir de que $\mathbb{E}(e_t|x_t) = 0$. Por lo tanto, el proyector lineal no necesariamente será una buena aproximación de la esperanza condicional.

2.2 Mínimos cuadrados ordinarios (OLS)

2.2.1 Preliminares

Vamos a comenzar con una muestra, que denotaremos por $\{(y_t, x_t) : t = 1, \ldots, T\}$ donde $y_t \in \mathbb{R}$ y $x_t \in \mathbb{R}^k$ vamos a asumir de que estas observaciones vienen de una distribución común (también se dice homogénea). Por lo tanto, un supuesto importante será el siguiente:

- Las variables $\{(y_1, x_1), \ldots, (y_T, x_T)\}$ están idénticamente distribuida y por ende, vienen de una distribución común F.

Nosotros lo que deseamos, es obtener un estimador para el coeficiente de proyección: $\beta = \mathbb{E}(x_t'x_t)^{-1}\mathbb{E}(x_t'y_t)$. Antes de derivar OLS, vamos a introducir el métodos de momentos:

Suponga que se desea estimar la media poblacional de una variable Y con distribución F:

$$\mu = E(Y) = \int_{-\infty}^{\infty} y dF(Y) \tag{31}$$

Un estimador natural es la media muestra:

$$\hat{\mu} = \bar{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t \tag{32}$$

Es así como el métodos de momentos, lo que busca es reemplazar la esperanza[3] por el promedio.

2.2.2 Derivando OLS

Lo que nos interesa encontrar es una norma que minimize la distancia entre la variable endógena y las variables exógenas, el método OLS lo que busca es minimizar el error del modelo al cuadrado. Otra manera de entenderlo, es que OLS es un estimador del coeficiente de proyección lineal, usando el métodos de momentos pero omitiendo el divido en T.
Teniendo esto presente, podemos volver a denotar a la suma de cuadrado de los residuos de la forma [4]:

$$S_T(\beta) = u'u = (Y - X\beta)'(Y - X\beta) = \underbrace{Y'Y - 2Y'X\beta + \beta'X'X\beta}_{\text{Es una matriz } 1\times 1 \text{ es decir un vector}} \tag{33}$$

Graficando dicha función en un espacio de $\mathbb{R}^3$, tendremos lo siguiente:

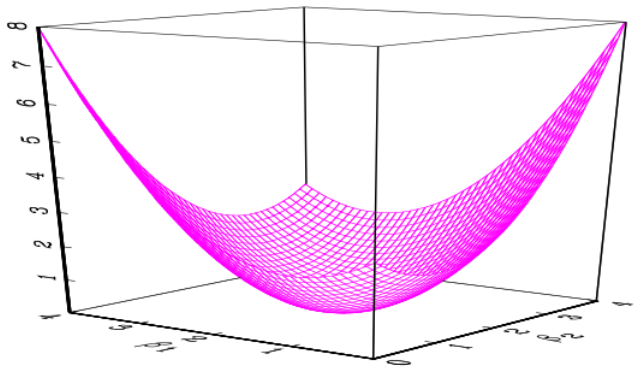


Figure 3.1: Sum-of-Squared Errors Function

OLS lo que busca es minimizar beta de S_T, por lo tanto:

$$\min_{\{\beta\}} S_T = \min_{\{\beta\}} u'u = \min_{\{\beta\}} (Y - X\beta)'(Y - X\beta) \tag{34}$$

La CPO será de la forma:

$$\frac{\partial S_T(\beta)}{\partial \beta}|_{\beta=\hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0 \longrightarrow X'X\hat{\beta} = X'Y \tag{35}$$

Por el supuesto de rango completo en las X, tendremos lo siguiente :

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{36}$$

[3]Esto es debido a que la esperanza es desconocida, puesto de que debo conocer todos los parámetros de la distribución F, los cuales no tengo

[4]Ocuparemos en lo que sigue notación matricial.

Para comprobar de que es un mínimo, podemos sacar la segunda derivada:

$$\frac{\partial^2 S_T(\beta)}{\partial\beta\partial\beta}|_{\hat{\beta}} = 2X'X \tag{37}$$

Definición: Definimos que una matriz B es semi definida positiva, si para todo $c \neq 0$, $c'Bc \geq 0$. En la medida que la matriz se pueda escribir $D'D$, siempre va a ser semi definida positiva:

$$\underbrace{c'D'}_{\alpha'}\underbrace{Dc}_{\alpha} = \alpha'\alpha = \sum \alpha_t^2 \geq 0 \tag{38}$$

Por lo tanto, $X'X$ siempre va a ser semi definida positiva y si le agregamos que es de rango completo es invertible [5] y, por ende, es definida positiva. Encontrando que efectivamente es un mínimo.A continuación, nombraremos unas particularidades de los coeficientes β:

- Es una función lineal de Y.
- Es una variable aleatoria, debido a que es una función de X e Y.
- Muestras distintas pueden presentar un coeficiente de correlación distinto.
- OLS se puede entender como un estimador del coeficiente de proyección lineal, el cual usa el métodos de momentos.

A continuación introduciremos dos matrices, la primera denominaremos matriz proyección:

$$P = X(X'X)^{-1}X' \tag{39}$$

La particularidad de esta matriz, es que es idempotente y simétrica, es decir todos sus valores propios son 0 o 1 ,en este caso la matriz P es de dimensiones $T \times T$, la cual presenta rango k con T eigenvalues dentro de los cuales K son 1 y T-K son 0, la idempotencia se puede ver con claridad, de la siguiente forma:

$$P \cdot P' = X\underbrace{(X'X)^{-1}X'X}_{I}(X'X)^{-1}X' = X(X'X)^{-1}X' = P \tag{40}$$

Por otro lado, la simetría se puede demostrar dado de que:

$$P' = \left[X(X'X)^{-1}X'\right]' = X(X'X)^{-1}X' = P \tag{41}$$

Denominamos a la matriz de proyección ortogonal (o complemento), de la forma:

$$M = I - X(X'X)^{-1}X' = I - P \tag{42}$$

La matriz M es una matriz $T \times T$, la cual es idempotente y simétrica, además presenta T-K valores propios igual a 1 y K valores propios igual a 0. La forma de demostrar idempotencia es la siguiente:

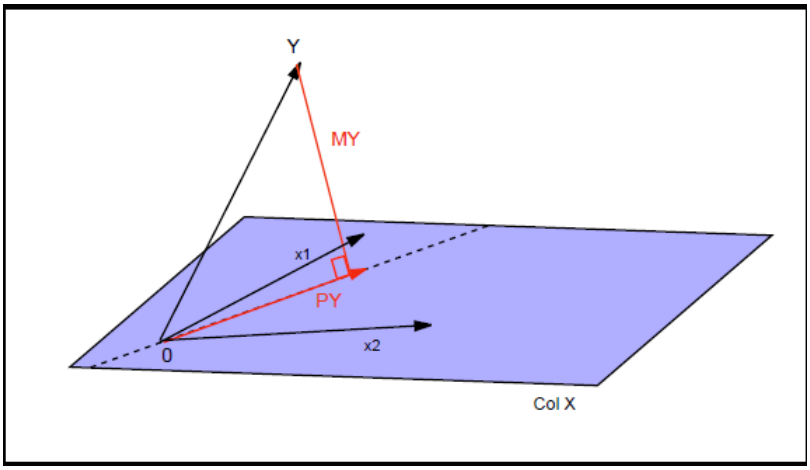
$$M \cdot M' = (I-P)'(I-P) = I' \cdot I'I - I'P - P'I + P'P = I - 2P + P = I - P = M \tag{43}$$

Con estas dos matrices, podemos recrear la descomposición ortogonal de Y, como sigue a continuación:

$$Y = MY + PY \tag{44}$$

Ols me garantiza de que la matriz MY sea ortogonal por construcción al espacio generado por los X (PY), como se puede observar a continuación:

[5] Una forma de que no sea invertibles es que los elementos dentro de ella presenten un grado de dependencia lineal. Es decir, que algún vector dentro de dicha matriz sea transformación lineal de algún otro, en el mismo espacio.



Lo que hace OLS, es que me descompone el Y, en una matriz que es explicada por los X's (PY) y otra matriz que es explicada por los $\hat{u}$ (MY) y no por las X's. Esto me asegura de que los X sean ortogonales a $\hat{u}$[6] pero no necesariamente al u del modelo real.
A continuación, haremos un poco de matemáticas con la matriz proyección y la matriz de proyección ortogonal, para lo cual tendremos presente la siguiente relación:

$$\hat{u} = Y - X\hat{\beta} \quad e \quad \hat{Y} = X\hat{\beta} \tag{45}$$

Llegammos a la siguiente relación:

$$PY = X(X'X)^{-1}X'Y = X(X'X)^{-1}X'(X\hat{\beta} + \hat{u}) = X\hat{\beta} + X(X'X)^{-1}X'\hat{u} = X\hat{\beta} = \hat{Y} \tag{46}$$

Esto es debido a que, por construcción de OLS $X'\hat{u} = 0$. Un caso especial ocurre cuando X = 1, es decir es un T-vector de unos, por lo tanto:

$$P_1 = 1(1'1)^{-1}1' = \frac{1}{T}11' \tag{47}$$

Por lo tanto:

$$P_1 Y = 1(1'1)^{-1}1'Y = 1\bar{Y} \tag{48}$$

Se crea en un T-vector cuyo elementos son las medias de y_i. Es sencillo demostrar que los eigenvalues de la matriz van entre 0 y 1, para lo cual usaremos la traza que se compone de la suma de los eigenvalues para demostrar lo siguiente:

$$\sum_{i=1}^{T} h_{ii} = tr(P) = K \quad \text{Por lo tanto,} \quad 0 \leq h_{ii} \leq 1 \tag{49}$$

Demostración:

$$tr(P) = tr(X(X'X)^{-1}X') = tr((X'X)^{-1}(X'X)) = tr(I_K) = K \tag{50}$$

Trabajando con la matriz de proyección ortogonal:

$$MY = Y - PY = Y - X\hat{\beta} = \hat{u} \tag{51}$$

El nombre de dicha matriz se debe a la propiedad de que para cualquier matriz Z perteneciente al rango de X, se cumple de que:

$$MZ = Z - PZ = 0 \tag{52}$$

Es relativamente sencillo, demostrar de que:

$$tr(M) = tr(I - P) = tr(I) - tr(P) = T - K \tag{53}$$

[6] Es decir $X'u = 0$.

Al igual que antes, podemos ver el caso de que cuando X=1, tendremos de que $P_1 = 1(1'1)^{-1}1'$ y por lo tanto:

$$M_1 Y = I_T Y - P_1 Y = I_T Y - 1\bar{Y} \tag{54}$$

Por lo tanto, podemos escribir el lado como $Y - \bar{Y}$, es decir el elemento i'th es $y_i - \bar{Y}$, el valor expresado en desviaciones de la media. Podemos usar la matriz M, para descomponer $\hat{e}$ con la siguiente relación:

$$\hat{u} = MY = (I - P)Y = Y - PY \tag{55}$$

Por lo visto en la ecuación 46, tendremos de que $PY = X\hat{\beta}$, por lo tanto:

$$\hat{u} = MY = Y - PY = Y - X\hat{\beta} = \hat{u} \tag{56}$$

Demostrando así la relación, usando esta misma podemos demostrar lo siguiente:

$$\hat{u} = MY = M(X\beta + u) = \underbrace{MX}_{=0}\beta + Mu = Mu \tag{57}$$

Por lo tanto, $\hat{u}$ está libre de ser dependiende del coeficiente de regresión β.
Usando la ecuación 45 y las ecuaciones 46 y 51, llegamos a la descomposición ortogonal antes nombrada:

$$Y = MY + PY \tag{58}$$

2.2.3 Varianza y Esperanza

A continuación, veremos los conceptos de esperanza y varianza para el modelo más básico de regresión lineal, con todos los supuestos antes nombrados. Para lo cual, usaremos el beta de OLS, derivado anteriormente, el cual es:

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + u) = \beta + (X'X)^{-1}X'u \tag{59}$$

Por lo tanto:

$$\hat{\beta} - \beta = (X'X)^{-1}X'u \tag{60}$$

Teniendo esto presente, denotaremos a la esperanza de la forma[7]:

$$\mathbb{E}\left[(\mathbb{E}(\beta) - \beta)|X\right] = \mathbb{E}\left[(\hat{\beta} - \beta)|X\right] = \mathbb{E}\left[(X'X)^{-1}X'u|X\right] \tag{61}$$

Si los X son determinísticos y no estocástico, llegaremos a:

$$\mathbb{E}\left[(\mathbb{E}(\beta) - \beta)|X\right] = \mathbb{E}\left[(X'X)^{-1}X'u|X\right] = (X'X)^{-1}X'\underbrace{\mathbb{E}\left[u|X\right]}_{=0} = 0 \tag{62}$$

Usando esperanza iteradas, tendremos de que:

$$\mathbb{E}(\mathbb{E}\left[(\mathbb{E}(\beta) - \beta)|X\right]) = \mathbb{E}(\hat{\beta} - \beta) = \mathbb{E}(0) = 0 \tag{63}$$

Por lo tanto, bajo los supuestos antes nombrados el estimador OLS es insesgado. No obstante, insesgamiento es un supuesto poco realista e infantil, debido a que la esperanza no es más que ponerse en distintos escenarios, sacar el promedio de cada uno y luego sacar el promedio de dichos promedio, por ende pocas veces los datos serán insesgado, más adelante veremos de que el supuesto que vamos a pedirle al estimador es el de eficiencia asintótica.
Para derivar la varianza, asumiremos (por el momento) de que hay homocedasticidad de los errores. Teniendo esto presente:

$$V(\hat{\beta}|X) = \mathbb{E}\left[(\mathbb{E}(\beta) - \beta)(\mathbb{E}(\beta) - \beta)'|X\right] = \mathbb{E}\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X\right] \tag{64}$$

[7]En este caso usaremos la Ley simple de esperanza iterada, la cual nos dice de que $\mathbb{E}(\hat{\beta}) = \mathbb{E}(\mathbb{E}(\beta|X))$

$$= \mathbb{E}\left[\left((X'X)^{-1}X'u\right)\left((X'X)^{-1}X'u\right)'|X\right] = \mathbb{E}\left[(X'X)^{-1}X'uu'X(X'X)^{-1}|X\right] \tag{65}$$

$$= (X'X)^{-1}X'\mathbb{E}\left[uu'|X\right]X(X'X)^{-1} = (X'X)^{-1}X'\sigma_u X(X'X)^{-1} \tag{66}$$

Dado de que hay homocedasticidad σ_u es un escalar, por lo tanto:

$$V(\hat{\beta}|X) = \sigma_u(X'X)^{-1}\underbrace{X'X(X'X)^{-1}}_{=I} = \sigma_u(X'X)^{-1} \tag{67}$$

Por lo tanto:

$$V(\hat{\beta}) = \mathbb{E}(V(\beta|X)) + V(\mathbb{E}(\hat{\beta}|x)) = \mathbb{E}(V(\beta|X)) + \underbrace{V(\beta)}_{=0} = \mathbb{E}(V(\beta|X)) = \sigma_u\mathbb{E}((X'X)^{-1}) \tag{68}$$

En el caso que los X's sean estocástico queda así, en el caso de que los X's sean determinístico la esperanza de los X es lo mismo que las X, por lo tanto el caso determinístico queda de la forma:

$$V(\hat{\beta}) = V(\hat{\beta}|X) = \sigma_u(X'X)^{-1} \tag{69}$$

Existen importantes características de la varianza:

- Crece proporcionalmente con σ_u.
- Decrece con el tamaño de la muestra (Ya que $X'X$ se hace más grande), por lo tanto a mayor tamaño es menor la variación y por ende convergerá al verdadero β.
- Decrece con la volatilidad de las variables exógenas.

Algo importante, es notar de que bajo estos supuestos el error medio cuadrático es igual a la varianza, esto debido a que por contrucción:

$$MSE = Var + Sesgo^2 \tag{70}$$

Como el sesgo es 0, tendremos de que:

$$MSE = Var \tag{71}$$

2.2.4 Aproximación de la varianza de los errores

El error de varianza es de la forma $\sigma^2 = \mathbb{E}u'u$, si u fuera observado, entonces podríamos estimar σ^2 por:

$$\tilde{\sigma}^2 = \frac{1}{T}u'u \tag{72}$$

Sin embargo, esto no se cumple si u no es observado. En este caso, hacemos dos pasos , el primero es estimar el modelo y obtener los residuos $\hat{u}$, luego reemplazar esto en la última ecuación, con lo que llegamos:

$$\hat{\sigma}^2 = \frac{1}{T}\hat{u}'\hat{u} \tag{73}$$

Como tenemos el hecho de que $\hat{u} = MY = Mu$, tendremos de que:

$$\hat{\sigma}^2 = \frac{1}{T}\hat{u}'\hat{u} = \frac{1}{T}Y'MMY = \frac{1}{T}Y'MY = \frac{1}{T}u'Mu \tag{74}$$

Una implicancia interesante, es que:

$$\tilde{\sigma}^2 - \hat{\sigma}^2 = \frac{1}{T}u'u - \frac{1}{T}u'Mu = \frac{1}{T}u'\left(I - M\right)u = \frac{1}{T}u'\left(I - I + P\right)u = \frac{1}{T}u'Pu \geq 0 \tag{75}$$

Esto debido a que la matriz proyección es semi-definida positiva y $u'Pu$ es una forma cuadrática. La relación anterior nos dice de que $\hat{\sigma}^2$ es numéricamente menor a $\tilde{\sigma}^2$. Si sacamos las esperanza de ambos estimadores tendremos lo siguiente:

$$\mathbb{E}(\hat{\sigma}^2|X) = \mathbb{E}(T^{-1}u'Mu|X) = T^{-1}\mathbb{E}(\underbrace{u'Mu}_{1\times 1}|X) = T^{-1}\mathbb{E}(tr(u'Mu)|X) = T^{-1}\mathbb{E}(tr(Muu')|X) \tag{76}$$

$$= T^{-1}tr(M\mathbb{E}(uu'|X)) = T^{-1}tr(M\sigma^2 I) = T^{-1}\sigma^2 tr(M) = \sigma^2(T-K)T^{-1} \tag{77}$$

Aplicando la Ley de esperanza iteradas, tendremos de que:

$$\mathbb{E}(\hat{\sigma}^2) = \sigma^2(T-K)T^{-1} \tag{78}$$

Se puede observar de que este estimador de la varianza es insesgado. Este estimador típicamente se utiliza en muestras grandes, no obsten en muestras pequeñas se tiende a usar el siguiente estimador insesgado:

$$\tilde{\sigma}^2 = \sigma^2(T-K)^{-1} \tag{79}$$

Es trivial demostrar de que bajo el modelo de regresión normal la varianza es la siguiente:

$$V(\hat{\sigma}^2) = T^{-2}2(T-K)\sigma^4 \tag{80}$$

La normalidad, solo funciona para demostrar este último resultado, pero para todo lo demás no es requerida, además tendremos los siguientes puntos a considerar:

- $\hat{\sigma}^2$ es sesgada, pero es de mínima varianza bajo el supuesto de normalidad (MLE) y es consistente.
- La varianza de $\hat{\beta}$ y de $\hat{\sigma}^2$ dependen de σ.

2.2.5 Análisis de varianza (ANOVA)

A continuación derivaremos la descomposición de la varianza de una estimación, para lo cual tendremos presente que la descomposición ortogonal es:

$$Y = PY + MY = \hat{Y} + \hat{u} \tag{81}$$

Por lo tanto [8]:

$$(Y-\bar{Y})'(Y-\bar{Y}) = (\hat{Y}-\bar{Y})'(\hat{Y}-\bar{Y}) + 2(\hat{Y}-\bar{Y})'\hat{u} + \hat{u}'\hat{u} \tag{82}$$

Por definición tendremos lo siguiente:

$$\hat{Y}'\hat{u} = (PY)'(MY) = Y'PMY = Y'(P(I-P))Y = Y'\underbrace{(P - \overbrace{P'P}^{P})}_{0}Y = 0 \tag{83}$$

Para eliminar el término $\bar{Y}'\hat{u}$ sea igual 0, tendremos lo siguiente:

$$\bar{Y}'\hat{u} = \bar{Y}\,\wr'\,\hat{u} \tag{84}$$

En donde $\wr$ es un vector de unos de dimensión T, por lo tanto teniendo presente de que $\wr'\hat{u}$ es el promedio de los u's es 0, siempre y cuando la regresión incluya constante. De esta manera, teniendo estas dos implicancia llegamos a:

$$\underbrace{(Y-\bar{Y})'(Y-\bar{Y})}_{\text{Total sum square (TSS)}} = \underbrace{(\hat{Y}-\bar{Y})'(\hat{Y}-\bar{Y})}_{\text{Explained sum square (ESS)}} + \underbrace{\hat{u}'\hat{u}}_{\text{Residual sum square (SSR)}} \tag{85}$$

[8] Teniendo presente de que $Y - \bar{Y} = \hat{Y} + \hat{u} - \bar{Y}$ por descomposición ortogonal.

Esto es típicamente llamado la fórmula de análisis de la varianza para la regresión de mínimos cuadrados. Un estadístico comúnmente reportada es el coeficiente de determinación o también llamado R-square (R cuadrado), el cual se formula de la siguiente manera:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} = 1 - \frac{\hat{u}'\hat{u}}{Y'LY} = 1 - \frac{Y'MY}{Y'LY} \tag{86}$$

Teniendo en mente de que $L = I_T - T^{-1}\wr\wr'$. Llegar al numerador es trivial, llegar al denominador lo derviaremos a continuación:

$$(Y - \bar{Y})'(Y - \bar{Y}) = Y'Y - Y'\bar{Y} - \bar{Y}'Y + \bar{Y}'\bar{Y} = Y'Y - 2Y'\bar{Y} + \bar{Y}'\bar{Y} \tag{87}$$

$$\bar{Y} = \begin{pmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix} = \begin{pmatrix} T^{-1}\sum_t y_t \\ \vdots \\ T^{-1}\sum_t y_t \end{pmatrix} = T^{-1}\begin{pmatrix} \sum_t y_t \\ \vdots \\ \sum_t y_t \end{pmatrix} \tag{88}$$

$$\wr' Y = (1 \dots 1)\begin{pmatrix} y_1 \\ \vdots \\ y_t \end{pmatrix} = \sum_t y_t \tag{89}$$

Sustituyendo en $\bar{Y}$ tenemos,

$$\hat{Y} = T^{-1}\begin{pmatrix} \wr'Y \\ \vdots \\ \wr'Y \end{pmatrix} = T^{-1}\begin{pmatrix} \wr \\ \vdots \\ \wr \end{pmatrix}\wr' Y = T^{-1}\wr\wr'Y \tag{90}$$

Por lo tanto:

$$Y'Y - 2Y'\bar{Y} + \bar{Y}'\bar{Y} = Y'Y - 2T^{-1}Y'\wr\wr'Y + (T^{-1}\wr\wr'Y)'(T^{-1}\wr\wr'Y) \tag{91}$$

$$= Y'Y - 2T^{-1}Y'\wr\wr'Y + T^{-1}Y'\wr\underbrace{(\wr'\wr)}_{=T}\wr' YT^{-1} \tag{92}$$

$$= Y'Y - 2T^{-1}Y'\wr\wr'Y + T^{-1}Y'\wr\wr'Y = Y'Y - T^{-1}Y'\wr\wr'Y = Y'\underbrace{(I_T - T^{-1}\wr\wr')}_{L}Y \tag{93}$$

Haciendo que se cumpla:

$$(Y - \bar{Y})'(Y - \bar{Y}) = Y'LY \tag{94}$$

Comprobando la ecuación:

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{Y'MY}{Y'LY} \tag{95}$$

Esta última relación nos ayuda a entender qué porcentaje de la suma de cuadrados está representado por el modelo. Hay varias consideraciones que tener en mente:

- Por construcción $0 \leq R^2 \leq 1$, si toma valor de 0 significa de que no hay nada en el componente sistemático que me ayude a predecir el comportamiento del Y. Sin embargo, uno tiende a pensar (erróneamente) de que se espera de que R^2 tome valor a 1, no obstante uno no espera que su modelo sea una definición o una tautología, ya que por definición es una aproximación a algún fenómeno y no el marco en si mismo.
- En términos conceptuales, no explica nada puesto de que es la economía la que explica los fenómenos. Es más, el R^2 no permite discriminar si un modelo es mejor que otro, debido a que carece de interpretación alguna, más adelante veremos criterios de selección de modelos que nos ayuda a escoger qué modelo es mejor que otro.

- Si el modelo no incluye constante, puede darse de que el R^2 sea negativo, ya que la regresión podría tener peor ajuste que la media.
- Un gran problema del R^2, es que al agregar regresores la SSR decrece o se mantiene constante, mientras que la TSS permanece constante.

Es por esta última razón de que nace el R^2 ajustado[9], el cual crece y decae por incluir más regresores:

$$\bar{R}^2 = 1 - \frac{SSR}{TSS} \cdot \frac{T}{T-K} = 1 - \frac{\tilde{\sigma}^2}{\hat{\sigma}_y^2} \tag{96}$$

Como se puede ver, el lado derecho se compone del estimador insesgado de la varianza de los $u's$ y del estimador sesgado de la varianza de y. Como se puede observar, el R^2 ajustado puede caer si el aporte del regresor es bajo. Teniendo esto definido, no nos queda más que decir de que en la actualidad se usa otros criterios de selección de modelos, los cuales veremos más adelante.

2.3 Estimador OLS de una partición

Supongamos de que tenemos la siguiente partición:

$$X = [X_1 \quad X_2] \quad \text{y} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \tag{97}$$

Podemos escribir el modelo de regresión lineal como:

$$Y = X_1\beta_1 + X_2\beta_2 + u \tag{98}$$

Podemos descomponer la operación $X'X\hat{\beta} = X'Y$ como:

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{bmatrix} X_1' \\ X_2' \end{bmatrix} Y \tag{99}$$

$$X_1'X_1\hat{\beta}_1 + X_1'X_2\hat{\beta}_2 = X_1'Y \tag{100}$$

$$X_2'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 = X_2'Y \tag{101}$$

De la ecuación 101, tendremos de que:

$$X_2'X_2\hat{\beta}_2 = X_2'Y - X_2'X_1\hat{\beta}_1 \quad /(X_2'X_2)^{-1}\cdot \tag{102}$$

$$\longleftrightarrow \hat{\beta}_2 = (X_2'X_2)^{-1}(X_2'Y - X_2'X_1\hat{\beta}_1) \tag{103}$$

Reemplazando esta relación en 100, tendremos de que:

$$X_1'X_1\hat{\beta}_1 + X_1'X_2(X_2'X_2)^{-1}(X_2'Y - X_2'X_1\hat{\beta}_1) = X_1'Y \tag{104}$$

$$= X_1'X_1\hat{\beta}_1 + X_1'X_2(X_2'X_2)^{-1}X_2'Y - X_1'X_2(X_2'X_2)^{-1}X_2'X_1\hat{\beta}_1 = X_1'Y \tag{105}$$

$$= X_1'X_1\hat{\beta}_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1\hat{\beta}_1 = X_1'Y - X_1'X_2(X_2'X_2)^{-1}X_2'Y \tag{106}$$

$$= X_1'(I - X_2(X_2'X_2)^{-1}X_2')X_1\hat{\beta}_1 = X_1'(I - X_2(X_2'X_2)^{-1}X_2')Y \tag{107}$$

[9]Este es el R^2 de Thail, hay otros pero este es el que más se usa.

Teniendo presente que $M_i = I - P_i$ y que $P_i = X_i(X_i'X_i)^{-1}X_i'$ para todo $i = 1, 2$, llegamos a:

$$X_1'M_2X_1\hat{\beta}_1 = X_1'M_2Y \tag{108}$$

$$\therefore \quad \hat{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2Y \tag{109}$$

El resultado es trivial para $\hat{\beta}_2$:

$$\therefore \quad \hat{\beta}_2 = (X_2'M_1X_2)^{-1}X_2'M_1Y \tag{110}$$

Esta es una aplicación de Frisch-Waugh-Lovell: $\hat{\beta}_2$ y $\hat{u}$ puede ser computarizado usando el siguiente comando:

1. Regresionando Y en X_1, obtenemos el residual $\tilde{Y}$.
2. Regresionamos X_2 en X_1, obtenemos el residual $\tilde{X}$.
3. Regresionamos $\tilde{Y}$ en $\tilde{X}$ y obtenemos $\hat{\beta}$ y el residual $\hat{u}$.

Este teorema se puede usar en el caso particionado, para lo cual tomaremos:

$$\hat{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2Y \tag{111}$$

Podemos denotar a esta expresión de la forma:

$$\hat{\beta}_1 = (X_1'^*X_1^*)^{-1}X_1'^*Y^* \tag{112}$$

En donde $X_1^* = M_2X_1$ e $Y = M_2Y$. En otras palabras X_1^* es el residuo de una regresión de X_1 en X_2 e Y^* es el residuo de la regresión de Y en X_2, por lo tanto por definición es el Teorema de Frisch-Waugh-Lovell. Es importante decir, de que dado de que se sacan los errores de las dos etapadas, para luego regresionarlos uno con el otro, se omite la constante, ya que en este caso la esperanza de los u son iguales a 0, por lo que no tiene sentido de que la pendiente no pase por otro lado que no sea el origen. Adicionalmente, Cabe destacar dos cosas:

- Para sacar la partición, el R^2 y el ajustado, solo se necesitan los supuestos de que las variables estén idénticamente distribuidas y que X'X sea de rango completo.
- Para sacar la esperanza insesgada y la varianza, se necesitaron los demás supuestos ya nombrados en un principio.

2.3.1 Teorema de Gauss-Markov

El Teorema de Gauss-Markov nos dice que bajo los supuestos del modelo de regresión lineal homocedástico, OLS es el mejor estimador lineal insesgado (se le dice que es Melli o Blue)[10], para demostrar esto denotaremos una matriz $A = (X'X)^{-1}X'$, por lo tanto $\hat{\beta} = AY$, tomemos cualquier otro estimador $\tilde{\beta} = (A + C)Y$.Por lo tanto:

$$\tilde{\beta} = (A + C)Y = (A + C)(X\beta + u) = (A + C)X\beta + (A + C)u = \underbrace{AX\beta}_{=\beta} + CX\beta + (A + C)u \tag{113}$$

$$\longrightarrow \mathbb{E}(\tilde{\beta}|x) = \mathbb{E}\left(\beta + CX\beta + (A + C)u|X\right) = \mathbb{E}\left(\beta|X\right) + \mathbb{E}(CX\beta|X) + \underbrace{\mathbb{E}\left((A + C)u|X\right)}_{=0} \tag{114}$$

$$\therefore \quad \mathbb{E}(\tilde{\beta}|x) - \beta = CX\beta \tag{115}$$

[10] Es importante esto último, dado que es el mejor estimador, pero comparándolo solo con estimadores que cumplan con esa condición.

Para que pertenezca a la familia de los estimadores insesgados se tiene que cumplir de que $CX\beta = 0$. Supongamos que se cumple esto, es sencillo ver de que:

$$\tilde{\beta} = (A+C)Y = \beta + \underbrace{CX\beta}_{=0} + (A+C)u = \beta + (A+C)u \tag{116}$$

Por lo tanto la varianza será de la forma:

$$V(\tilde{\beta}|X) = \mathbb{E}\left[(\tilde{\beta} - \mathbb{E}(\tilde{\beta}))(\tilde{\beta} - \mathbb{E}(\tilde{\beta}))'|X\right] = \mathbb{E}\left[((A+C)u)((A+C)u)'|X\right] \tag{117}$$

$$= \mathbb{E}\left[(A+C)uu'(A+C)'|X\right] = (A+C)\underbrace{\mathbb{E}\left[uu'|X\right]}_{=\sigma^2}(A+C)' = \sigma^2(A+C)(A+C)' \tag{118}$$

Es sencillo descomponer $(A+C)(A+C)'$ de la forma:

$$AA' + 2CA' + CC' = (X'X)^{-1}\underbrace{X'X(X'X)^{-1}}_{=I} + 2\underbrace{CX}_{=0}(X'X)^{-1} + CC' \tag{119}$$

$$\therefore \quad (A+C)(A+C)' = (X'X)^{-1} + CC' \tag{120}$$

Juntando esta última expresión con 118 llegamos a lo siguiente:

$$V(\tilde{\beta}|X) = \sigma^2((X'X)^{-1} + CC') = \sigma^2(X'X)^{-1} + \sigma^2 CC' \tag{121}$$

Como podemos recordar, la varianza de OLS era la siguiente:

$$V(\hat{\beta}|X) = \sigma^2(X'X)^{-1} \tag{122}$$

Así, nos queda lo siguiente:

$$V(\tilde{\beta}|X) = V(\hat{\beta}|X) + \sigma CC' \longrightarrow V(\tilde{\beta}|X) - V(\hat{\beta}|X) = \sigma^2 CC' \tag{123}$$

A medida de que CC' sea semi-definida postiva se va a cumplir siempre de que:

$$V(\tilde{\beta}|X) \geq V(\hat{\beta}|X) \tag{124}$$

Si no se cumpliera homocedasticidad OLS dejará de ser el mejor estimador lineal insesgado. Otro punto interesante a evaluar, es que se cumple de que también es el que tiene menor error cuadrático medio, ya que en este caso como es insesgado el ECM será igual a la varianza.

2.4 Mínimos cuadrados restringidos (CLS)

A continuación, veremos una aplicación de mínimos cuadrados la cual se basa en la idea de que enfrentamos reestricciones al valor de los coeficientes β. Supongamos que se tiene que cumplir la siguiente reestricción:

$$Q'\beta = C$$

Donde Q es una matriz de $(K \times q)$ constantes conocidas y C es una matriz de q-vectores constantes conocidos, las cuales cumplen las siguientes condiciones:

- El rando de Q es q, donde $q < K$ (se debe cumplir con desigualdad estricta).
- Existen q reestricciones.

Teniendo esto en mente, diremos que el estimador CLS, es tal que minimiza la SSR sujeto a $Q'\beta = C$, por lo tanto podemos hacer uso de nuestro Lagrangiano como se muestra a continuación [11]:

$$\mathcal{L}(\beta,\gamma) = \underbrace{\underbrace{(Y-X\beta)'(Y-X\beta)}_{OLS} + 2\gamma'(Q'\beta - C)}_{CLS} \tag{125}$$

Donde γ es de dimensión $1 \times q$ y es un $q-vector$ del multiplicador de Lagrange, tendremos que las FONC serán del estilo:

$$\frac{\partial \mathcal{L}}{\partial \beta}|_{\bar{\beta},\bar{\gamma}} = 0 \longrightarrow -2X'Y + 2X'X\bar{\beta} + 2Q\bar{\gamma} = 0 \tag{126}$$

$$\frac{\partial \mathcal{L}}{\partial \gamma}|_{\bar{\beta},\bar{\gamma}} = 0 \longrightarrow Q'\hat{\beta} = C \tag{127}$$

Para encontrar los coeficientes, multiplicaremos la ecuación 126 por $Q'(X'X)^{-1}$:

$$-2Q'\underbrace{(X'X)^{-1}X'Y}_{\hat{\beta}} + 2\underbrace{Q'\underbrace{(X'X)^{-1}X'X}_{I}\bar{\beta}}_{Q'\bar{\beta}=C} + 2Q'(X'X)^{-1}Q\bar{\gamma} = 0 \tag{128}$$

$$-Q'\hat{\beta} + C + Q'(X'X)^{-1}Q\bar{\gamma} = 0 \longleftrightarrow Q'(X'X)^{-1}Q\bar{\gamma} = Q'\hat{\beta} - C \quad / \quad \left[Q'(X'X)^{-1}Q\right]^{-1}\cdot \tag{129}$$

$$\therefore \quad \bar{\gamma} = \left[Q'(X'X)^{-1}Q\right]^{-1}(Q'\hat{\beta} - C) \tag{130}$$

Reemplazando esta expresión en 126, tendremos lo siguiente:

$$X'X\bar{\beta} = X'Y - Q\left[Q'(X'X)^{-1}Q\right]^{-1}(Q'\hat{\beta} - C) \quad / \quad (X'X)^{-1}\cdot \tag{131}$$

$$= \bar{\beta} = \underbrace{(X'X)^{-1}X'Y}_{=\hat{\beta}} - (X'X)^{-1}Q\left[Q'(X'X)^{-1}Q\right]^{-1}(Q'\hat{\beta} - C) \tag{132}$$

$$\therefore \quad \bar{\beta} = \hat{\beta} - (X'X)^{-1}Q\left[Q'(X'X)^{-1}Q\right]^{-1}(Q'\hat{\beta} - C) \tag{133}$$

Por definición $\hat{\beta} = (X'X)^{-1}X'Y$ e $Y = X\beta + u$, por lo tanto $\hat{\beta} = \beta + (X'X)^{-1}X'u$, reemplazando este expresión en 133 llegamos a:

$$\therefore \quad \bar{\beta} = \beta + (X'X)^{-1}X'u - (X'X)^{-1}Q\left[Q'(X'X)^{-1}Q\right]^{-1}(Q'\beta + Q'(X'X)^{-1}X'u - C) \tag{134}$$

Teniendo presente la segunda condición de optimalidad ($Q'\beta = C$), llegamos a:

$$\therefore \quad \bar{\beta} = \beta + (X'X)^{-1}X'u - (X'X)^{-1}Q\left[Q'(X'X)^{-1}Q\right]^{-1}Q'(X'X)^{-1}X'u \tag{135}$$

Llegando así al estimador CLS, a continuación derivaremos la esperanza y la varianza:

$$\mathbb{E}(\bar{\beta}|X) = \beta + (X'X)^{-1}X'\underbrace{\mathbb{E}(u|X)}_{=0} - (X'X)^{-1}Q\left[Q'(X'X)^{-1}Q\right]^{-1}Q'(X'X)^{-1}X'\underbrace{\mathbb{E}(u|X)}_{=0} \tag{136}$$

$$\therefore \quad \mathbb{E}(\bar{\beta}) = \beta \tag{137}$$

Es insesgado, por otro lado si definimos:

$$A = I - Q\left[Q'(X'X)^{-1}Q\right]^{-1}Q' \tag{138}$$

[11] Al multiplicador le agregaremos un dos por conveniencia, pero sin el 2 los resultados no cambian.

$$V(\bar{\beta}) = \mathbb{E}\left[(\bar{\beta}-\beta)(\bar{\beta}-\beta)'\right] = \mathbb{E}\left[(A(X'X)^{-1}X'u)(A(X'X)^{-1}X'u)'\right] \tag{139}$$

$$= \mathbb{E}\left[A(X'X)^{-1}X'uu'X(X'X)^{-1}A'\right] = A(X'X)^{-1}X'\mathbb{E}(uu)'X(X'X)^{-1}A' \tag{140}$$

$$\therefore \quad V(\bar{\beta}) = \sigma^2 A(X'X)^{-1}A' \tag{141}$$

Se puede demostrar de que lamatriz A es idempotente, pero no simétrica. En este caso, el estimador de la varianza será de la forma:

$$\bar{\sigma}^2 = T^{-1}(Y - X\bar{\beta})'(Y - X\bar{\beta}) \tag{142}$$

Esta varianza será mayor o igual al $\hat{\sigma}$.

2.5 Inferencia

El modelo normal de regresión lineal como vimos en un principio, agrega el supuesto de normalidad de los errores:

$$u \sim N(0, \sigma I) \tag{143}$$

La ventaja de asumir normalidad, yace en la capacidad de poder hacer inferencia exacta, lo cual significa que los estadísticos tienen distribuciones conocidas incluso en muestras finitas. Bajo estos supuestos, dado de que $\hat{\beta}$ depende de la distribución de u, tendremos de que:

$$\hat{\beta}|X \overset{a}{\sim} N(\beta, \sigma^2(X'X)^{-1}) \tag{144}$$

Es importante, que es aproximada como normal. Por otro lado, es claro notar de que como $\hat{\beta} \xrightarrow{P} \beta$, esto además converge a una distribución degenerada (colapsa en un solo punto, que en este caso es beta), por lo que necesitaremos algo más de inferencia (con teoría asintótica lo veremos con más detalle).
Asimismo, cualquier combinación lineal de $\hat{\beta}$ tendrá distribución normal:

$$\hat{\theta}|X = Q'\hat{\beta} \overset{a}{\sim} N(\underbrace{Q'\beta}_{=C}, \sigma^2 Q'(X'X)^{-1}Q) \tag{145}$$

A continuación, discutiremos una muestra finita (exacta) y grande de estimadores para hacer test de hipótesis, cuyos componentes serán:

- Hipótesis nula: H_0
- Hipótesis alternativa: H_1

La idea de estos dos conceptos, es que nosotros tenemos una teoría que asumimos verdadera la cual nunca estamos seguro de ella, por lo tanto la contrarrestamos con los datos. Es importante acusar de que la hipótesis nula está fuertemente protegida, ya que le exigimos un nivel de significancia, pero nunca se acepta, solo que no se rechaza. Esto último es fundamental en las ciencias, ya que sabemos de que nuestro modelo es solo una aproximación de la estructura real. Además, tenemos dos tipos de test:

- El test estadístico será de una cola o de dos colas,en el primer tipo me interesa el signo y en la me interesa la distancia, además me interesa con igualdad o desigualdad.
- Teniendo estos elementos, tendremos una zona de rechazo y una conclusión.

Usando la inferencia con constraste lineal y usando el supuesto normales, testearemos la siguiente hipótesis:

$$H_0 : Q'\beta = C \quad \wedge \quad H_1 : Q'\beta \neq C \tag{146}$$

Con $q = 1$ ($Q'\beta$ escalar), asumiendo normalidad, bajo la hipótesis nula:
Con σ conocido:

$$Z = \frac{Q'\hat{\beta} - C}{[\sigma^2 Q'(X'X)^{-1}Q]^{\frac{1}{2}}} \sim N(0,1) \tag{147}$$

Rechazaremos la hipótesis nula a un nivel de significancia α si:

$$|Z| < Z_{1-\alpha/2}$$

Donde $Z_{1-\alpha/2}$ se obtiene de la tabla de distribución normal (0,1).
Cuando σ es desconocido, usamos el estadístico (seguimos asumiendo normalidad)[12]:

$$\frac{\hat{u}'\hat{u}}{\sigma^2} \sim \chi^2_{T-K} \tag{148}$$

Adicionalmente, usando el hecho de que:

$$\frac{N(0,1)}{\sqrt{\chi^2/(T-K)}} \tag{149}$$

Tendremos de que con σ desconocido y q=1:

$$t_T = \frac{Q'\hat{\beta} - C}{\sqrt{\tilde{\sigma} Q'(X'X)^{-1}Q}} \sim S_{T-K} \tag{150}$$

Es decir se convierte en un t-student, la cual tiene la particularidad de ser simétricas pero con kurtosis mayor que la distribución normal, además tendremos de que cuando $T \longrightarrow \infty$ se convierte en una normal. Además rechazaremos la hipótesis nula a un nivel de significancia α, si es que:

$$|t_T| > S_{1-\alpha/2} \tag{151}$$

Donde $S_{1-\alpha/2}$ se obtiene de la tabla de distribución de t-student. Para construir intervalos de confianza, tendremos de que:

$$1 - \alpha = Pr\left[S_{[\alpha/2,T-K]} \leq \frac{Q'\hat{\beta} - C}{\sqrt{\tilde{\sigma} Q'(X'X)^{-1}Q}} \leq S_{[1-\alpha/2,T-K]}\right] \tag{152}$$

$$\longleftrightarrow 1 - \alpha = Pr\left[Q'\hat{\beta} - \sqrt{\tilde{\sigma} Q'(X'X)^{-1}Q} S_{[1-\alpha/2,T-K]} \leq Q'\beta \leq Q'\hat{\beta} - \sqrt{\tilde{\sigma} Q'(X'X)^{-1}Q} S_{\alpha/2,T-K}\right] \tag{153}$$

Por lo tanto:

$$Q'\beta \in \left[Q'\hat{\beta} \pm \sqrt{\tilde{\sigma} Q'(X'X)^{-1}Q} S_{[1-\alpha/2,T-K]}\right] \tag{154}$$

A continuación, estudiaremos el caso en que $q > 1$, es decir, cuando $Q'\beta$ es una matriz:
Cuando σ es conocido:[13]

$$(Q'\hat{\beta} - C)'(\sigma^2 Q'(X'X)^{-1}Q)^{-1}(Q'\hat{\beta} - C)' \sim \chi^2_q \tag{155}$$

Cuando σ es desconocido: [14]

$$\frac{\frac{(Q'\hat{\beta}-C)'(\sigma^2 Q'(X'X)^{-1}Q)^{-1}(Q'\hat{\beta}-C)'}{q}}{\frac{\frac{\hat{u}'\hat{u}}{\sigma^2}}{T-K}} \sim F_{q,T-K} \tag{156}$$

$$\frac{S_T(\bar{\beta}) - S_T(\hat{\beta})}{\tilde{\sigma}^2} = \frac{T-K}{q} \frac{(Q'\hat{\beta} - C)'[Q'(X'X)^{-1}Q]^{-1}(Q'\hat{\beta} - C)}{\hat{u}'\hat{u}} \sim F_{q,T-K} \tag{157}$$

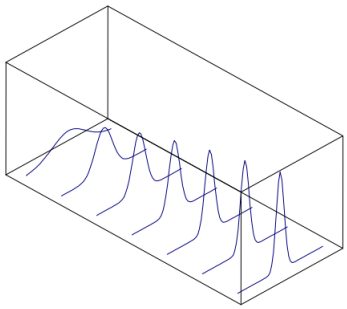
En el caso de más de una reestricción, para general intervalos y regiones de confianza.

$$Pr\left[(Q'\hat{\beta} - Q'\beta)'(\tilde{\sigma} Q'(X'X)^{-1}Q)^{-1}(Q'\hat{\beta} - Q'\beta)/q \leq F_{q,T-K}\right] = 1 - \alpha \tag{158}$$

[12] Esto debido a que en el numerador tenemos la multiplicación de dos normales estándar lo que distribuye como chi cuadrado con un grado de libertad.
[13] Esto es debido, a que como argumentamos anteriormente la multiplicación de dos normales estándar, dan como resultado una χ^2_1, por lo tanto la suma de q chis da como resultado una χ^2_q.
[14] La distribución F se origina de dos χ^2 divididas por sus respectivos grados de libertad.

2.6 Teoría asintótica

A continuación, como hacer inferencia cuando u no es necesariamente normal, como discutimos con anterioridad a medida que aumento mi tamaño de muestra la matriz de varianza-covarianza se vuelve más pequeña y el coeficiente $\hat{\beta}$ converge al parámetro real, como se muestra a continuación:



La teoría asintótica es la propiedad de los estimadores cuando el tamaño de muestra es infinitamente grande (o muy grande), por lo tanto hay contexto en los cuales las propiedades de muestras pequeñas y grandes se asemejan, es por esto último que necesitamos saber qué tan complejo es el problema para decidir si apoyarnos en una o en otra. A continuación, hablaremos de las conocidas por la teoría y a las usadas para estos estimadores:

- Convergencia en probabiliad: Diremos de que una secuencia de vectores reales o de variables aleatorias $\{x_t\}$ converge al parámetros real x en probabilidad si:

$$\lim_{T \longrightarrow \infty} Pr\left(\|x_t - x\| > \mathbb{E}\right) = 0 \quad \forall \mathbb{E} > 0 \tag{159}$$

  Donde $\|x_t - x\|$ es una norma, típicamente se utiliza la norma eucledіana. Notros a esto lo denotamos como $x_t \xrightarrow{P} x$ o $Plim \;\; x_t = x$.

- Convergencia en el cuadrado de las medias: Se dice que $\{x_t\}$ converge en medias cuadradas si:

$$\lim_{T \longrightarrow \infty} \mathbb{E}(x_t - x)^2 = 0 \tag{160}$$

  Nosotros la escribimos de la forma $x_t \xrightarrow{M} x$.

- Convergencia casi segura: $\{x_t\}$ converge casi seguro a x si:

$$Pr\left[\lim_{T \longrightarrow \infty} x_t = x\right] = 1 \tag{161}$$

  En este caso se denota como $x_t \xrightarrow{a.s} x$. Es importante, acusar de que $a.s \longrightarrow M \longrightarrow P$.

- El estimador $\hat{\theta}_T$ de θ_0 es llamado estimador débilmente consistente si $\hat{\theta}_T \xrightarrow{P} \theta_0$.
- El estimador $\hat{\theta}_T$ de θ_0 es llamado estimador fuértemente consistente si $\hat{\theta}_T \xrightarrow{a.s} \theta_0$.
- Teorema WLLN1 (Ley débil 1) Chebyshev : Teniendo $\mathbb{E}(x_t) = \mu_t$, $v(x_t) = \sigma_t$ y $Cov(x_t, x_j) = 0 \;\; \forall i \;\; \neq j$, tendremos de que si:[15]

$$\lim_{T \longrightarrow \infty} \frac{1}{T} \sum_{t=1}^{\infty} \sigma_t \leq M \leq \infty \tag{162}$$

[15] Es importante, que este teorema deja que los x cambien con los errores, además permite que la varianza sea heterocedástica, puesto a que puede variar según la muestra (notar los sub-índices).

Es decir, si dicha expresión convege a un parámetros finito, se cumple de que:

$$\bar{x}_T - \bar{\mu}_T \xrightarrow{P} 0 \tag{163}$$

Donde $\bar{x}_T$ y $\bar{\mu}_T$ representan el promedio muestra y el promedio de la esperanza muestral de μ_t respectivamente.

- SLLN1 (Ley fuerte) Kolmogorov : $\{x_t\}$ es idependiente con varianza finita $v(x_t) = \sigma_t^2 < \infty$. Si:

$$\sum_{t=1}^{\infty} \frac{\sigma_t^2}{t^2} < \infty \tag{164}$$

Entonces:

$$\bar{x}_T - \bar{\mu}_T \xrightarrow{a.s} 0 \tag{165}$$

Esto es relevante, ya que nos explica el por qué el supuesto de independencia es un supuesto más fuerte que suponer que las covarianzas son iguales a 0.

Para OLS, tendremos de que si:

$$T^{-1}X'X \longrightarrow Q \quad \forall \quad \text{Q invertible y no estocástico} \tag{166}$$

Entonces:

$$\hat{\beta} - \beta = (X'X)^{-1}X'u \;/\cdot \frac{T}{T} \tag{167}$$

$$\therefore \underbrace{(T^{-1}X'X)^{-1}}_{=Q^{-1}}(T^{-1}X'u) \xrightarrow{P} 0 \tag{168}$$

Así, diremos de que el estimador $\hat{\beta}$ es (débilmente) consistente, ya que:

$$\hat{\beta} \xrightarrow{P} \beta \tag{169}$$

Cabe destacar, de que la consistencia es un hecho más relevante (y realista) que el sesgo.

- Convergencia en distribución: $\{x_t\}$ se le dice que converge a x en distribución, si la distribución $\mathcal{F}_t$ de x_t converge a $\mathcal{F}$ de x en toda la continuidad de puntos de $\mathcal{F}$, se escribe $x_T \xrightarrow{D} x$ y llamamos a $\mathcal{F}$ como el límite de la distribución de $\{x_t\}$. Si $\{x_t\}$ e $\{y_t\}$ tienen el mismo límite de distribución, lo denotamos como $x_T \overset{LD}{=} y_T$.
- CTL1 (Teorema central del límite) Lindeberg-Lévy: Teniendo a $\{x_t\}$ i.i.d. con $\mathbb{E}x_t = \mu$ y $Vx_t = \sigma^2$, entonces: [16]

$$Z_t = \frac{\bar{x}_T - \mu}{[V\bar{x}_T]^{1/2}} = \sqrt{T}\frac{\bar{x}_T - \mu}{\sigma} \xrightarrow{D} N(0,1) \tag{170}$$

Donde $\bar{x}_T$ representa el promedio muestral de los x.

Llevando esto último a OLS, tendremos de que asumiendo $T^{-1}X'X \longrightarrow Q$ invertible y no estocástico, y $T^{-1/2}X'u \xrightarrow{D} N(0, \sigma^2 Q)$, entonces:

$$\sqrt{T}(\hat{\beta} - \beta) = (T^{-1}X'X)^{-1}(T^{-1/2}X'u) \xrightarrow{D} N(0, \sigma^2 Q^{-1}) \tag{171}$$

Por lo tanto, bajo el modelo homocedástico la distribución asintótica no depende de la distribución de los errores.

[16] Nótese que incluimos el supuesto de que los x son independientes e identicamentes distribuidos, lo cual es un supuesto más fuerte que no correlación, además todas las muestras son iguales independiente de T, finalmente no hay presencia de heterocedasticidad.

2.7 Test de quiebre estructural

Supongamos de que tenemos 2 régimenes de regresión:

$$Y_1 = X_1\beta_1 + u_1 \tag{172}$$

$$Y_2 = X_2\beta_2 + u_2 \tag{173}$$

Los betas son distintos para ambas regiones, teniendo de que $T_1 + T_2 = T$, además nos interesa la varianza σ^2 antes y después del evento:

$$\mathbb{E}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}\begin{bmatrix} u_1' u_2' \end{bmatrix} = \begin{bmatrix} \sigma_1^2 I_{T_1} & 0 \\ 0 & \sigma_2^2 I_{T_2} \end{bmatrix} \tag{174}$$

En le modelo reestringido , la hipótesis nula es la siguiente[17]:

$$H_0 : \beta_1 = \beta_2 \tag{175}$$

Nuestra hipótesis nula nos dice de que no hubo cambios antes y después del evento, asumiendo de que $\sigma_1 = \sigma_2$ (es decir que tienen la misma volatilidad):

$$Y = \underline{X}\beta + u \tag{176}$$

Donde:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}, \quad \underline{X} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \quad \text{y} \quad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{177}$$

Aplicando:

$$\frac{S_T(\bar{\beta}) - S_T(\hat{\beta})}{\tilde{\sigma}^2} = \frac{T-K}{q}\frac{(Q'\hat{\beta} - C)'[Q'(X'X)^{-1}Q]^{-1}(Q'\hat{\beta} - C)}{\hat{u}'\hat{u}} \sim F_{q,T-K} \tag{178}$$

Llegamos a:

$$\frac{T_1 + T_2 - 2K}{K} \cdot \frac{(\hat{\beta}_1 - \hat{\beta}_2)'[(X_1'X_1)^{-1} + (X_2'X_2)^{-1}]^{-1}(\hat{\beta}_1 - \hat{\beta}_2)}{Y'[I - \underline{X}(\underline{X}'\underline{X})^{-1}\underline{X}']Y} \sim F_{K,T_1+T_2-2K} \tag{179}$$

Donde $\hat{\beta}_i = (X_i'X_i)^{-1}X_iY \quad \forall \quad i = 1, 2.$
El mismo resultado puede ser derivado a continuación: Si definimos la SSR bajo la alternativa (cambios estructurales):

$$S_T(\hat{\beta}) = Y'[I - \underline{X}(\underline{X}'\underline{X})^{-1}\underline{X}']Y \tag{180}$$

Y SSR bajo la hipótesis nula:

$$S_T(\bar{\beta}) = Y'[I - X(X'X)^{-1}X']Y \tag{181}$$

Es fácil mostrar de que:

$$\frac{T_1 + T_2 - 2K}{K} \cdot \frac{S_T(\bar{\beta}) - S_T(\hat{\beta})}{S_T(\hat{\beta})} \sim F_{K,T_1+T_2-2K} \tag{182}$$

En este caso un estimador insesgado de la varianza es:

$$\tilde{\sigma}^2 = \frac{S_T(\bar{\beta})}{T_1 + T_2 - 2K} \tag{183}$$

Antes, nosotros habíamos removido el supuesto de que $\sigma_1 = \sigma_2$, nosotros primeros derivaremos un test de igual varianza. Bajo la hipótesis nula (igual varianza a través de los regímenes), tendremos de que:

$$\frac{\hat{u}_i'\hat{u}_i}{\sigma^2} \sim \chi^2_{T_i-K} \quad \forall \quad i = 1, 2. \tag{184}$$

[17] En el modelo no reestringido incluyo toda la muestra.

Dado que las chi-cuadradas son independientes, tendremos de que:

$$\frac{T_2 - K}{T_1 - K} \cdot \frac{\hat{u}_1' \hat{u}_1}{\hat{u}_2' \hat{u}_2} \sim F_{T_1-K, T_2-K} \tag{185}$$

A diferencia de las pruebas anteriores, aquí debe utilizarse una prueba de dos colas, porque un valor grande o pequeño valor de la prueba es una razón para rechazar la hipótesis nula. Si eliminamos la hipótesis de igualdad de varianzas entre regímenes y nos centramos en la hipótesis de igualdad de los parámetros de regresión, las pruebas son más complicadas. Nos concentraremos, en el caso cuando K=1, aquí un t-test es aplicable, se puede demostrar de que:

$$t_T = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\frac{\tilde{\sigma}_1}{X_1' X_1} + \frac{\tilde{\sigma}_2}{X_2' X_2}}} \sim S_v \tag{186}$$

Donde:

$$v = \frac{\left[\frac{\tilde{\sigma}_1}{X_1' X_1} + \frac{\tilde{\sigma}_2}{X_2' X_2}\right]^2}{\frac{\tilde{\sigma_1}^2}{(T_1-1)(X_1' X_1)^2} + \frac{\tilde{\sigma_2}^2}{(T_2-1)(X_2' X_2)^2}} \tag{187}$$

Aunque son muy usados los test de cambios estructurales (o Chow test), la econometría moderna es escéptica, ya que en estos test el econometrista establece una forma ad-hoc en cómo dividir la muestra. Las aplicaciones teóricas y empíricas recientes trabajan en tratar el periodo de posible ruptura como una variable latente endógena.

2.8 Boopstrap y Montecarlos

(Pendiente)

3 Formas funcionales

A continuación, varemos algunas extensiones de OLS las cuales son cruciales en la práctica.

3.1 Efectos de escalados

Los datos no siempre están convenientemente escalados, por ejemplo el peso de una persona puede estar determinado en gramos, kilogramos e inclusive toneladas, por lo tanto es esencial entender cuando las variables endógenas/exógenas cambian la cuantía.

3.1.1 Cambiando la escala de X

Supongamos que transformamos el siguiente modelo:

$$y_t = \beta_1 + \beta_2 x_t + u_t \tag{188}$$

Escalamos por c de la forma:

$$y_t = \beta_1 + c \cdot \beta_2 \frac{x_t}{c} + u_t \tag{189}$$

Con esto llegamos al siguiente modelo:

$$y_t = \beta_1 + \beta_2^* x_t^* + u_t \tag{190}$$

Lo que sucede en estos casos es que:

- Cambiamos la magnitud del coeficiente por c.
- Cambiamos el coeficiente del error estándar por el mismo factor.
- t-statics no es afectado.
- Todo lo demás se mantiene sin cambios.

3.1.2 Cambiando la escala de Y

Supongamos que tenemos el mismo modelo:

$$y_t = \beta_1 + \beta_2 x_t + u_t \tag{191}$$

Multiplicamos todo por $\frac{1}{c}$:

$$\frac{y_t}{c} = \frac{\beta_1}{c} + \frac{\beta_2}{c} x_t + \frac{u_t}{c} \tag{192}$$

Lo que es equivalente a:

$$y_t^* = \beta_1^* + \beta_2^* x_t^* + u_t^* \tag{193}$$

Lo que sucede en estos casos es que:

- Cambiamos todas las magnitudes por la constante.
- Cambiamos el coeficiente de las desviaciones estándar de los errores por el mismo factor.
- No afecta al t-test.
- Escalamos los residuales y cambiamos el error estándar de la regresión (SER) por el mismo factor.
- Todo lo demás se mantiene sin cambios.

3.2 Variables Dummys

En muchos casos nos interesa estudiar el efecto heterogéneo entre grupos, en estos casos podemos definir una Dummy siempre y cuando las variables que tomamos para comparar son independientes y exógenas, y tienen interpretación cualitativa. Por ejemplo, si definimos una Dummy de la forma:

$$d_{1,i} = \begin{cases} 1 & \text{Mujer} \\ 0 & \text{Hombre} \end{cases} \tag{194}$$

Por lo tanto, podemos escribir la ecuación de la forma:

$$w_i = \beta_0 + \beta_1 d_{1,i} + u_i \tag{195}$$

Donde los resultados serán:

- $\beta_0 = \mathbb{E}(wage|hombre) = \beta_0$.
- $\beta_0 = \mathbb{E}(wage|mujer) = \beta_0 + \beta_1$.

Donde β_1 es el castigo/premio por ser mujer. No obstante, nuestro análisis estaría sesgado ya que hay más factores que influyen en el salario de la persona, por lo tatno será un modelo insuficiente, una forma de representar esto es que $Cov(x, u) \neq 0$. Adicional a esto, podemos definir una segunda dummy de la forma:

$$d_{2,i} = \begin{cases} 0 & \text{Mujer} \\ 1 & \text{Hombre} \end{cases} \tag{196}$$

Y podemos estimar la siguiente regresión:

$$w_i = \beta_1 d_{1,i} + \beta_2 d_{2,i} + u_i \tag{197}$$

En este caso:

- $\beta_1 = \mathbb{E}(wage|mujer)$.
- $\beta_2 = \mathbb{E}(wage|hombre)$.

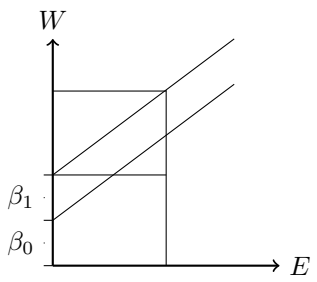
En estos casos para evaluar si no hay discriminación, podemos ver si la hipótesis $\beta_1 = \beta_2$ se cumple. No obstante, esta última forma no es muy usada, ya que nos basta encontrar que la primera es signifativa (la que incluye solo la dummy es mujer) para encontrar significancia.
Adicionalmente, podemos expandir nuestro análisis incluyendo más regresores:

$$w_i = \beta_0 + \beta_1 d_{1,i} + \beta_2 E_i + u_i \tag{198}$$

En este casos incluimos a la variable eduación, la cual puede ser cualitativa o cuantitativa, en este caso intercepto el efecto de los géneros, pero la educación es la misma, una manera de arreglar esto es:

$$w_i = \beta_0 + \beta_1 d_{1,i} + \beta_2 E_i + \beta_3 d_{1,i} E_i + u_i \tag{199}$$

Esta última ecuación, permite la diferencia de pendiente como se puede observar a continuación:



Como se puede observar, existe un premio por ser mujer, pero el retorno a la educación es el mismo porque tienen la misma pendiente las rectas. A medida que una pendiente es más inclinada que la otra, se argumenta de que existe un premio a la educación para el grupo que presente mayor inclinación.
Es interesante ver cómo nuestro estimador algebraicamente maneja las variables dummys:

$$W = D_1\beta_1 + D_2\beta_2 + u \tag{200}$$

Por construcción $D_1'D_2 = 0$, por lo tanto:

$$\hat{\beta}_1 = (D_1'D_1)^{-1}D_1'W = \frac{\sum_{i=1}^{N} d_{1,i}W_i}{\sum_{i=1}^{N} d_{1,i}W_i} = \frac{1}{N}\sum_{i=1}^{N} d_{1,i}W_i = \bar{W}_1 \tag{201}$$

Donde N_1 es el número de observaciones con $d_{1,i} = 1$ y $\bar{W}_1$ es la media muestral de estas observaciones. Paralelamente, tendremos de que:

$$\hat{\beta}_2 = \bar{W}_2 \tag{202}$$

Por lo tanto, para evaluar diferencias lo que necesito es hacer un teste de diferencias de medias, por lo cual necesito una matriz de varianzas-covarianzas de los betas, asumiendo que los ues son homocedásticos:

$$\hat{V}(\hat{\beta}) = \hat{\sigma}^2 \begin{bmatrix} (D_1'D_1)^{-1} & 0 \\ 0 & (D_2'D_2)^{-1} \end{bmatrix} = \begin{bmatrix} \frac{\hat{\sigma}^2}{N_1} & 0 \\ 0 & \frac{\hat{\sigma}^2}{N_2} \end{bmatrix} \tag{203}$$

Donde:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} \hat{u}_i^2 \tag{204}$$

Si hay heterocedasticidad:

$$\hat{V}(\hat{\beta}) = \begin{bmatrix} \hat{\sigma}_1^2(D_1'D_1)^{-1} & 0 \\ 0 & \hat{\sigma}_2^2(D_2'D_2)^{-1} \end{bmatrix} = \begin{bmatrix} \frac{\hat{\sigma}_1^2}{N_1} & 0 \\ 0 & \frac{\hat{\sigma}_2^2}{N_2} \end{bmatrix} \tag{205}$$

Donde:

$$\hat{\sigma}_j^2 = \frac{1}{N_j}\sum_{i=1}^{N} d_{j,i}\hat{u}_i^2 \quad \text{Para} \quad j = 1, 2. \tag{206}$$

3.3 Utilizando regresión particionada y dummys para crear efectos fijos

El efecto fijo, se basa en la idea de que uno sabe de que tiene efectos heterogéneo por ciertas variables, pero la cuantía de cuanto es el efecto es irrelevante. El caso más común es el de los colegios, por ejemplo si queremos estudiar cómo se relaciona una serie de regresores al puntaje de una prueba, es muy probable que tengamos que incluir una dummy por cada establecimiento educacional, no obstante el coeficiente particular de cada colegio es irrelevante para nuestro análisis, solo nos interesa saber cómo afectan al coeficiente de nuestras variables de interés. En estos casos, lo que podríamos hacer es utilizar regresión particionada de la forma:

$$Y = X_1\beta_1 + X_2\beta_2 + u \tag{207}$$

En donde X_1 contiene los regresores de los coeficientes que nos interesan conocer y X_2 contiene las dummys creadas para cada establecimiento (exceptuando uno que será nuestra base), por lo tanto según nuestro definición de regresión particionada, tendremos de que :

$$\hat{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2Y \tag{208}$$

De esta forma, obtenemos los regresores con el efecto fijo por establecimiento incluidos.

3.4 Tendencias de tiempo

Considere una serie que tiene una tasa de crecimiento constante:

$$z_t = z_0(1+g)^t \tag{209}$$

Es decir la variable z_t no tiene una relación lineal y presenta un crecimiento exponencial, si tomamos logaritmo:

$$\ln(z_t) = \ln(z_0) + t\ln(1+g) \tag{210}$$

Podemos cambiar la notación, agregándole un shock:

$$y_1 = \beta_1 + \beta_2 t + u_t \tag{211}$$

Donde t es la tendencia, β_2 es la tasa de crecimiento porcentual de g: $\beta_2 = \ln(1+g) \approx g$ (aproximadamente igual a g, cuando este es pequeño), independiente de la medida de t el beta no cambiará, pero si cambiará el estimador de la constante porque cambiará el intercepto.

3.4.1 Estacionalidad

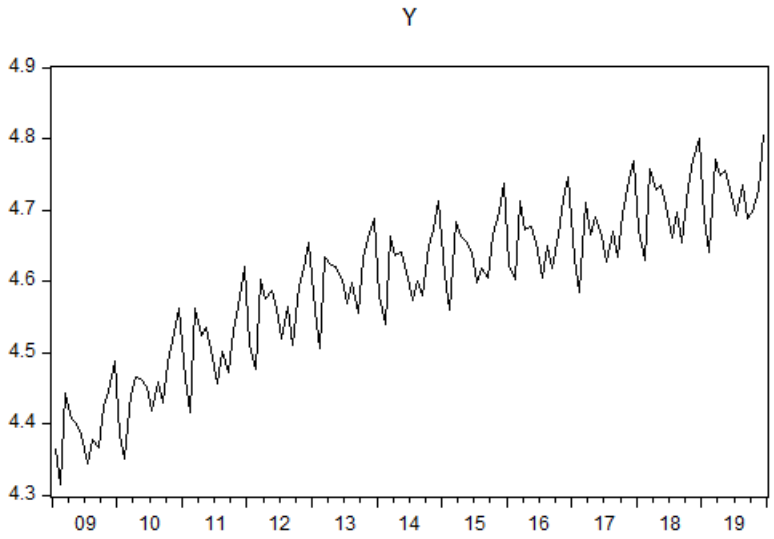
La idea básica es que las series de tiempo se pueden descomponer en 4 factores, los cuales son:

$$Y_t = S_t + C_t + T_t + U_t \tag{212}$$

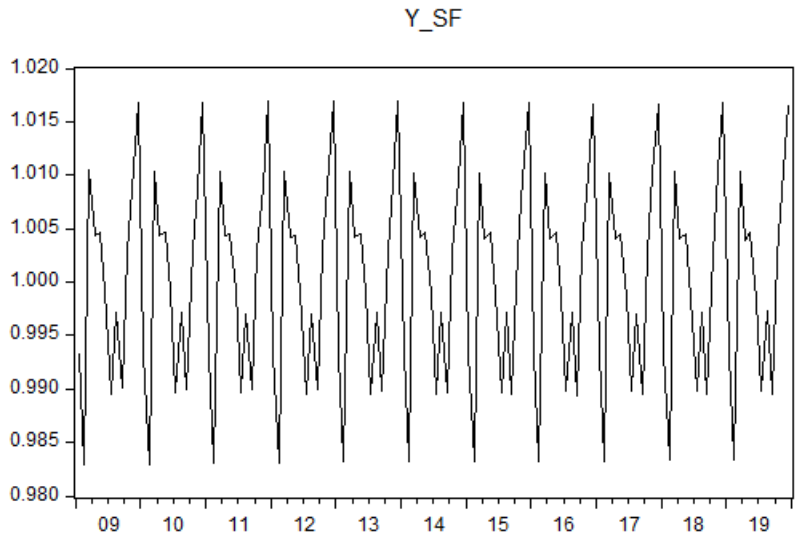
En donde:

- S_t es el factor estacional, el cual no tiene relación con la economía, es un componente que se repite de manera habitual y no tiene una explicación económica detrás (por lo general, se debe a tradiciones, costumbres como las fiestas patrias, etc.).
- C_t es el factor cíclico, es un componente que nos dice que hay un co-movimiento en el tiempo, cuyo signo depende de la correlación, ejemplo la persistencia de un boom.
- T_t es el factor tendencial, esta nos muestra si la serie presenta algún crecimiento de largo plazo.
- U_t es el factor irregular o no predecible.

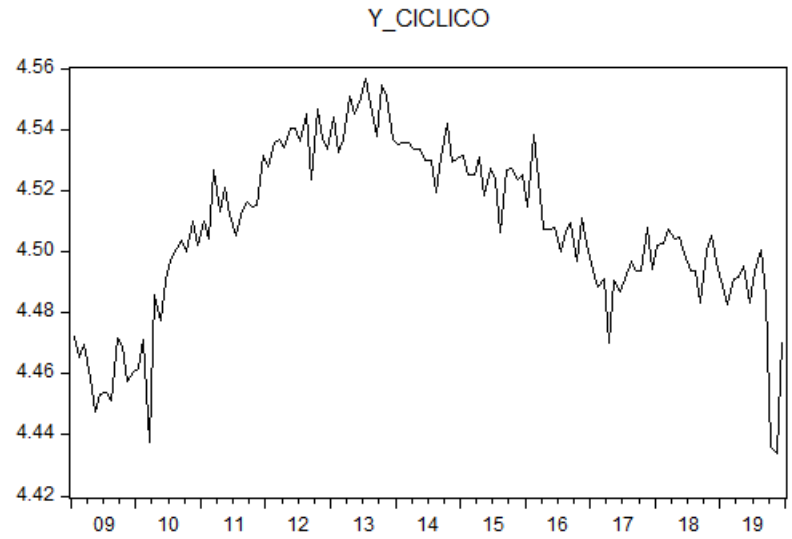
Supongamos que tenemos la siguiente serie:



Como se puede observar, hay un patró que se repite periodo tras periodo, este componente se denomina el estacional el cual se presenta a continuación:



El componente cíclico será de la siguiente forma:



La tendencia, es un tema que discutiremos en breve.
En economía, lo que nos interesa es desestacionalizar la serie, puesto de que este componente no nos interesa, para esto existen distintos métodos dependiendo de que si la serie presenta un comportamiento sistemático o aleatorio, partiremos con el primer caso. Supongamos de que los trimestres son distintos, lo que hacemos bajo este escenario, es estimar una regresión de la forma:

$$y_t = \beta_0 + \beta_1 q_{1,t} + \beta_2 q_{2,t} + \beta_3 q_{3,t} + \beta_4 t + u_t \tag{213}$$

Creamos tantas variables dummys como categorías, menos una si es que queremos incluir constantes. Para sacar el efecto estacional de la serie (desestacionalizar), hacemos lo siguiente con la serie:

$$y_t^* = y_t - (\hat{\beta}_1 q_{1,t} + \hat{\beta}_2 q_{2,t} + \hat{\beta}_3 q_{3,t}) \tag{214}$$

Así obtenemos la serie desestacionalizada de una manera determinística. Si nosotros tomamos la serie de manera aleatoria existen métodos llamados X12-ARIMA que desestacionaliza la serie con una serie de elementos que se escapan a los alcances de este apunte.

3.4.2 Calcular el componente tendencial de una serie

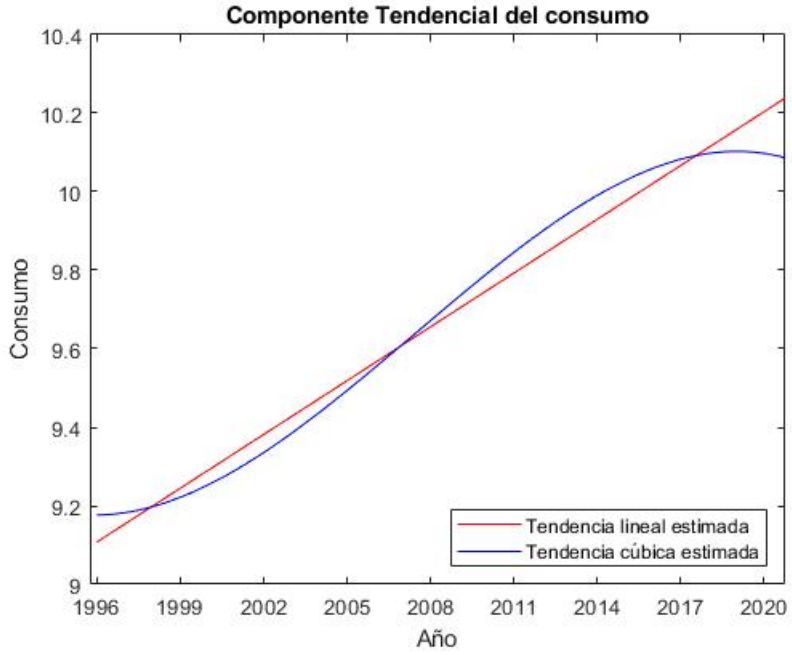
Lo que nos interesa en este inciso es sacar el componente tendencial, para lo cual hay dos formas de calcularlo, uno es suponiendo de que el crecimiento de la serie tiene un comportamiento determinístico y otro es suponer de que sigue un comportamiento aleatorio, para explicar esto supondremos que la serie con la que trabajamos ya se encuentra desestacionalizada, para calcular el componente tendencian de una manera determinística nosotros le agregamos la forma con que crece, usualmente se ocupa la forma lineal o cúbica (no cuadrática, por la construcción de estas), por lo tanto, hacemos una regresión de la siguiente forma:
Aproximación lineal, regresionamos:

$$y_t = \beta_0 + \beta_1 t + u_t \tag{215}$$

El componente $\beta_1 t$ será nuesto componente tendencial. Para la aproximación cúbica, regresionamos:

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + u_t \tag{216}$$

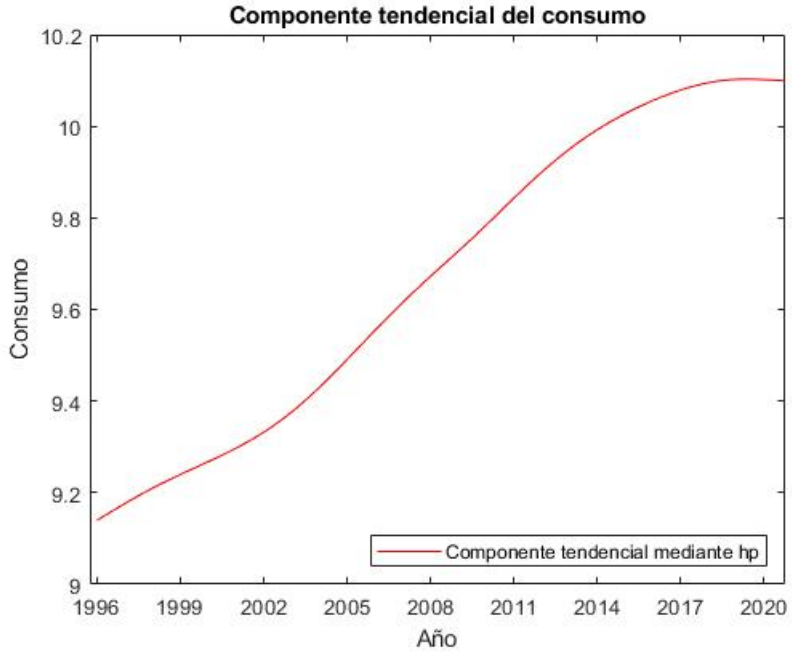
En este caso $\beta_1 t + \beta_2 t^2 + \beta_3 t^3$ es nuestra aproximación de la tendencia, con una tendencia cúbica. Gráficamente:



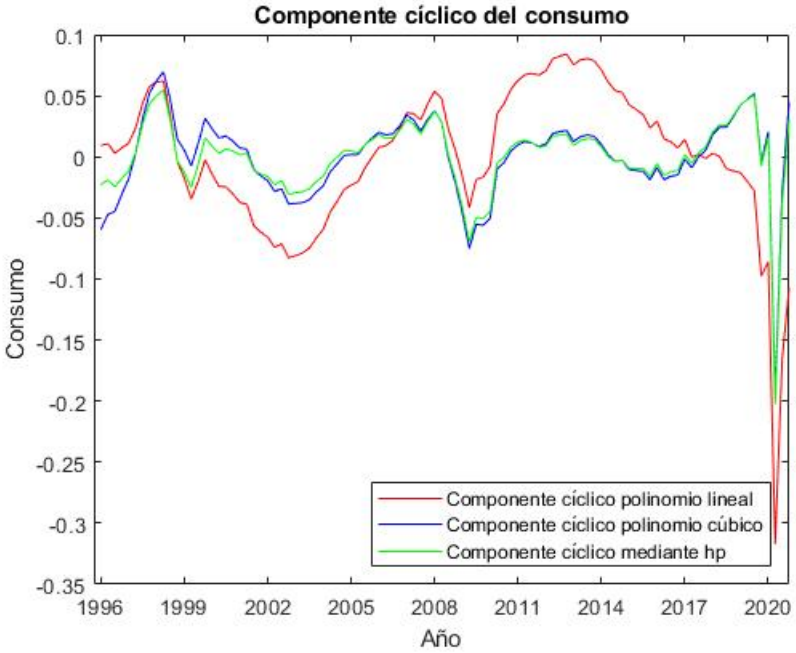
No obstante, típicamente las series macreconómicas presentan un crecimiento aleatorio, para estos casos típicamente se utiliza el filtro Hodrick-Prescott (Hodrick y Prescott, 1980), también llamado 'filtro HP', para el cual supondremos de que tenemos una serie de la forma $x_1, x_2, \ldots, x_T$ encontraremos $x_1^{tr}, x_2^{tr}, \ldots, x_T^{tr}$ que minimice:

$$\sum_{t=1}^{T}(x_t - x_t^{tr})^2 + \lambda \sum_{t=1}^{T-1}(\Delta x_{t+1}^{tr} - \Delta x_t^{tr})^2 \tag{217}$$

Este programa asume de que la tendencia se 'suaviza' por la penalización de la variación del segundo componente. El parámetro λ captura la importancia relativa de tener una tendencia suave versus el ajuste de la serie observada. Cuando λ crece la serie se vuelve más suave, en el extremo cuando $\lambda \longrightarrow \infty$, obtenemos una tendencia lineal, por el contrario cuando es 0 obtenemos una tendencia parecida a la serie en si. Además, los autores asumen de que un buen valor para este λ es 1600 [18], aplicando esto a nuestra serie llegamos a:



Para obtener el componente cícliclo, baste con restar la tendencia a la serie desestacionalizada llegando a lo siguiente:



[18] Idea profundizada por Kydland y Prescott (1990) y Wabha (1980).

Es claro notar que la varianza del componente cíclico cuando la aproximación de la tendencia es lineal es mayor, puesto de que la tendencia se ajusta menos a la serie.

3.5 Regresiones no lineales

Nosotros estamos interesados en la expresión de $\mathbb{E}(y_t|x_t) = m(x_t)\ \forall\ x \in \mathbb{R}$, donde la forma de m es desconocida. En OLS asumimos de que una buena aproximación es con coeficientes lineales, no obstante los x pueden ser de alguna otra forma. Podemos argumentar de que m es suficientemente suave para que se pueda aproximar a un polinomio de x:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \cdots + \beta_j x_t^j + u_t \tag{218}$$

Si $x \in \mathbb{R}^2$, una simple aproximación es:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{1,t}^2 + \beta_4 x_{2,t}^2 + \beta_5 x_{1,t} x_{2,t} + u_t \tag{219}$$

A medida de que la dimensión crece, la aproximación se hace no parsimoniosa. Si bien, la mayor parte usa términos cuadráticos, algunos utilizan cubos sin interacciones (redes neuronales, series de Fourer, etc). Esto puede ser o más eficiente o menos eficiente:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{1,t}^2 + \beta_4 x_{2,t}^2 + \beta_5 x_{1,t} x_{2,t} + \beta_6 x_{1,t}^3 + +\beta_7 x_{2,t}^3 + u_t \tag{220}$$

Como se puede observar, la regresión sigue siendo lineal en los coeficientes, la gran desventaja de la dimensionalidad, es que a medida que aumenta el problema se hace más complejo de manera exponencia (a esto se le denomina la maldición de la dimensionalidad).
Además, al incluir no linealidad la interpretación toma esto en cuenta. Por ejemplo, del modelo anterior:

$$\frac{\partial \mathbb{E}(y_t|x_t)}{\partial x_{1,t}} = \underbrace{\beta_1 + 2\beta_3 x_{1,t} + \beta_5 x_{2,t} + 3\beta_6 x_{1,t}^2}_{\text{Efecto no lineal y no constate}} \tag{221}$$

Como se puede observar, el efecto es una función de $x_{1,t}$ y $x_{2,t}$ lo que dificulta el reporte de la pendiente. Esta complejidad hace de que el efecto de la variable sobre otra, no sea constante y dependa del punto en el cual nos encontremos, para dar una solución a esto, típicamente la derivada se evalua en dos candidatos:
En la media muestral:

$$\frac{\partial \mathbb{E}(y_t|x_t)}{\partial x_{1,t}}|_{x_t=\bar{x}} = \beta_1 + 2\beta_3 \bar{x}_{1,t} + \beta_5 \bar{x}_{2,t} + 3\beta_6 \bar{x}_{1,t}^2 \tag{222}$$

En el promedio:

$$\frac{1}{T}\sum_{t=1}^{T} \frac{\partial \mathbb{E}(y_t|x_t)}{\partial x_{1,t}} = \beta_1 + 2\beta_3 \bar{x}_{1,t} + \beta_5 \bar{x}_{2,t} + 3\beta_6 \frac{1}{T}\sum_{t=1}^{T} x_{1,t}^2 \tag{223}$$

Donde el efecto evaluado en el promedio es distinto al efecto promedio en el componente cúbico($\mathbb{E}(F(x)) \neq F(\mathbb{E}(x))$).
Típicamente se escoge la segunda forma, ya que si x es muy simétrico el efecto promedio puede evaluar de mejor forma la diferencia.

3.5.1 Transformaciones

Muchos modelos simples consideran no linealidades, pero podemos usar transformaciones que nos da una aproximación bastante buena. Por ejemplo, usemos la función de producción Cobb-Douglas:

$$Q_i = U_i K_i^{\alpha} L_i^{\gamma} \tag{224}$$

Si tomamos logaritmo natural:

$$q_i = \alpha K_i + \gamma L_i + u_i \tag{225}$$

Quiero estimar α y γ, donde puedo usar OLS si K y u es ruido blanco al igual que L y u. Otro ejemplo, es la función:

$$Q_i = e^{x_i'\beta + u_i} \tag{226}$$

Al igual que antes, podemos tomar logaritmos naturales llegando a:

$$q_i = x_i'\beta + u_i \tag{227}$$

No obstante, hay funciones no lineales las cuales no se pueden aplicar transformaciones para transformarlas a lineales, un ejemplo claro es la función CES de la forma:

$$Q_i = \left[\alpha K_i^{\theta} + (1-\alpha)L_i^{\theta}\right]^{\frac{1}{\theta}} + u_i \tag{228}$$

En este último caso no se puede aplicar OLS.

Además, tenemos de que escoger la forma funcional de una relación afecta la interpretación de los resultados, acontinuación tenemos algunos casos típicos:

Nombre	Función	Derivada	Elasticidad
Lineal	$y = \beta_1 + \beta_2 x$	β_2	$\beta_2 \frac{x}{y}$
Cuadrática	$y = \beta_1 + \beta_2 x^2$	$2\beta_2$	$2\beta_2 \frac{x^2}{y}$
Cúbica	$y = \beta_1 + \beta_2 x^3$	$3\beta_2^2$	$3\beta_2^2 \frac{x^3}{y}$
log-log	$\ln y = \beta_1 + \beta_2 \ln x$	$\beta_2 \frac{y}{x}$	β_2
log-lineal	$\ln y = \beta_1 + \beta_2 x$	$\beta_2 y$	$\beta_2 x$
lineal-log	$y = \beta_1 + \beta_2 \ln x$	$\beta_2 \frac{1}{x}$	$\beta_2 \frac{1}{y}$

Los economistas tenemos dos preguntas que nos interesan responder, una es ¿qué ocurre con y cuando cambia x?, una primera respuesta es evaluar de pedniente (dy/dx) otras veces me interesa responder esta misma pregunta pero independiente de la escala de x e y, por lo que evaluamos la elasticidad $\Delta\%y/\Delta\%x \longleftrightarrow \frac{dy}{dx}\cdot\frac{x}{y}$, esta última expresión no depende de la unidad.

3.5.2 Ln(y) vs y como variable independiente

Los econometristas pueden estimar $Y = X\hat{\beta} + \hat{u}$ o $\ln Y = X\hat{\beta} + \hat{u}$ o ambas, pero ¿cuál es mejor?, la realidad es que cualquiera está bien a medida de que $\mathbb{E}(y_t|x_t)$ y $\mathbb{E}(\ln y_t|x_t)$ estén bien definidos ($Y > 0$).
Para seleccionar una sobre otra, necesitamos imponer otra estructura adicional, por ejemplo de que la esperanza sea lineal en x_t y $u \sim N(0, \sigma^2)$, en estos casos:
Razones para preferir $\ln(Y)$ sobre Y:

- Quizás $\mathbb{E}(\ln(y_t)|x_t)$ sea lineal en x_t, mientras de que $\mathbb{E}(y_t|x_t)$ puede ser no lineal, y por lo tanto tenemos que los modelos lineales son más sencillos de interpretar.
- $u_t = \ln(y_t) - \mathbb{E}(\ln(y_t)|x_t)$ los errores pueden ser menos heterocedásticos que los errores de una especificación lineal (al revés ocurre pocas veces).
- A medida de que $y_t > 0$, el rango en $\ln(y_t)$ está bien definida en los $\mathbb{R}$, esto no sucede en los casos en que $\hat{y}_t$ dado ciertos x_t y $\hat{\beta}$ puede producirse de que $\bar{y_t} < 0$ (Tobit).
- Si levantamos el supuesto de que u distribuye normal, puede darse de que la distribución sea asimétrica, por lo tanto $\mathbb{E}(y_t|x_t)$, no será una buena medida de tendencia central, y el estimador estará (será) influenciada por las observaciones en los extremos o "fuera de la linea" (se parece a una χ^2), por lo tanto $\mathbb{E}(\ln(y_t)|x_t)$ será una mejor medida de tendencia central [19].
- Hay que tener cuidado si se aplica logaritmo natural y se está interesado por obtener $\mathbb{E}(y_t|x_t)$, ya que se cumple desigualdad de Jensen:

$$\exp\left[\mathbb{E}(\ln(y_t)|x_t)\right] \neq \mathbb{E}\left[\exp(\ln(y_t))|x_t\right] \tag{229}$$

[19] Es importante decir de que, si a una distribución se le tiene que aplicar logaritmo para que distribuya normal, se dice que la serie es log-normal.

3.5.3 Estimadores para no-linealidades omitidas

Estos test son usados para cubrirnos de problemas de especificación, es decir para ver si estamos capturando de manera correcta la relación entre los x_t y los y_t. Una manera es agregar una función no lineal de x_t y testear significancia. Entonces podemos usar abuso de notación y ocupar $z_t = h(x_t)$ para denotar de esta forma una función no lineal de x_t, estimamos mediante OLS la siguiente expresión:

$$y_t = x_t'\tilde{\beta} + z_t'\tilde{\gamma} + \tilde{u}_t \tag{230}$$

Y testeamos $H_0 : \tilde{\gamma} = 0$.
Otro teste muy ocupado es el **Ramsey reset test** o también conocido como test de no linealidades omitidas, para este test partimos con una especificación simple en lugar de un modelo complejo:

$$y_t = x_t'\beta + u \tag{231}$$

Lo cual es estimado por OLS, obteniendo los valores predichos $\hat{y}_t = x_t'\hat{\beta}$. Ahora usaremos:

$$z_t = \begin{pmatrix} \hat{y}_t^2 \\ \vdots \\ \vdots \\ \hat{y}_t^j \end{pmatrix} \tag{232}$$

El cual es un vector $(j-1)$ de transformaciones de $\hat{y}_t^j$. Corriendo la regresión auxiliar:

$$y_t = x_t'\tilde{\beta} + z_t'\tilde{\gamma} + \tilde{u}_t \tag{233}$$

Por OLS y teniendo un estadístico wald W_T para $H_0 : \gamma = 0$, la hipótesis nula será de la forma $W_T \sim \chi^2_{j-1}$, esta nula es rechazada a un nivel $\alpha\%$ si el estadístico W_T excede por sobre $\alpha\%$ la cola crítica del valor de la distribución χ^2_{j-1}. Para hacer el test, j debe ser seleccionada con anterioridad. Típicamente, valores pequeños como j=2,3 o 4 dan mejores resultados.
El test RESET es una función suave, para detectar indicadores simples de la forma:

$$y_t = G(x_t'\beta) + u_t \tag{234}$$

Donde $G(\cdot)$ es una función suave de conexión, para ver el por qué la ecuación se puede escribir como:

$$y_t = x_t'\tilde{\beta} + (x_t'\tilde{\beta})^2\tilde{\gamma}_1 + (x_t'\tilde{\beta})^3\tilde{\gamma}_2 + \cdots + (x_t'\tilde{\beta})^j\tilde{\gamma}_{j-1} + \tilde{u}_t \tag{235}$$

Lo cual tiene la esencialidad de aproximar a $G(\cdot)$ por un polinomio de orden $j - th$. Es importante decir, de que heterocedasticidad es pariente de no linealidad, no obstante hay que ser cuidadoso respecto de esto último.

3.6 ¿Están los errores normalmente distribuidos?

Como vimos anteriormente, normalidad de los u's no es crucial para las propiedades del estimador OLS (incluyendo la inferencia), sin embargo conduce a una distribución exacta en muestras pequeñas, ayudando en pronóstico, etc. Por lo tanto, cada vez que tengamos una relación deberiamos hacer un test de nomalidad, a continuación veremos el test más típico, para lo cual necesitamos comentar de que necesitamos tan solo dos estalígrafos necesarios para definir una serie (primeros dos momentos, la media y la varianza), dado de que todo momento mayor va a ser una construcción de estos dos momentos.

3.6.1 Jarquer-Bera test

A continuación, veremos el test más común para testear normalidad, el cual se construye de la siguiente manera:

$$JB = \frac{T}{6} \cdot \left[S^2 + \frac{(K-3)^2}{4} \right] \xrightarrow{D} \chi^2_2 \tag{236}$$

Son dos grados de libertad, debido a que son dos hipótesis (S y K). Este test, se puede hacer para los residuos o para cualquier variable.

- S (asimetría o skewness en inglés) es una medida de asimetría de la distribución respecto de la media[20]:

$$S = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{z_t - \bar{z}}{S_z}\right)^3 = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\tilde{z}_t}{S_z}\right)^3 \tag{237}$$

  Cuando $S = 0$ es simétrica, en caso de que $S > 0$ la distribución tiene la cola ancha por la derecha, en caso de que $S < 0$ la distribución presenta cola ancha por la izquierda.

- K (kurtosis) es la medida de volumen o amplitud de la distribución, en otras palabras trata de explicar qué tan plana o parada es la distribución:

$$K = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{z_t - \bar{z}}{S_z}\right)^4 = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\tilde{z}_t}{S_z}\right)^4 \tag{238}$$

  Cuando K=3 la distribución es normal, con $K > 3$ la distribución presenta colas más anchas (Platykurtic) y en caso contraria, cuando $k < 3$ la distribución es presenta colas más planas que una distribución normal (Leptokurtic distribution).

3.7 Error de medición

A continuación veremos las consecuencias de hacer una mala medición, de manera no sistemática, para lo cual usaremos la siguiente regresión:

$$Y^* = X^*\beta + u \tag{239}$$

En donde X^* o Y^* no son observados, además tendremos de que $Y = Y^* + v$, $X = X^* + w$; $v \sim (0, \sigma_v^2 I)$, $w \sim (0, \sigma_w^2 I)$.
Considere primero de que Y está medido con error, en estos casos tendremos de que:

$$Y = X^*\beta + u + v = X^*\beta + u^*, \tag{240}$$

Donde $u^* = u + v$. Esto satisface el supuesto de LRM, por lo tanto $\hat{\beta}$ es insesgado y eficiente (pero menos que cuando y es observado), además tendremos de que

$$V(\hat{\beta}|X^*) = (\sigma_u^2 + \sigma_v^2)(X'X)^{-1}$$

Dado de que los errores no son sistemáticos y por ende $V(u + v|X^*) = (\sigma_u^2 + \sigma_v^2)I$, no obstante el estimador OLS sigue siendo MELI (BLUE).
Ahora supongamos de que X^* está medida con error, en estos casos:

$$Y = (X - w)\beta + u = X\beta + u^* \tag{241}$$

Donde $u^* = u - w\beta$, $X = X^* + w$, en estos casos los regresores estarán correlacionados con los errores, la correlación se puede derivar de la siguiente manera, para lo cual debemos recordar de que $\mathbb{E}(X|u) = 0$, es un supuesto más fuerte de que $\mathbb{E}(Xu) = 0$, por lo tanto basta que lo segundo no se cumpla para que lo primero tampoco, por lo tanto tendremos lo siguiente:

$$\mathbb{E}(Xu^*) = \mathbb{E}(X(u - w\beta)) = \mathbb{E}((X^* + w)(u - w\beta)) \tag{242}$$

$$= \mathbb{E}(X^*u + wu - X^*w'\beta - ww'\beta) = \underbrace{\mathbb{E}(X^*u)}_{=0} + \underbrace{\mathbb{E}(wu)}_{=0} - \underbrace{\mathbb{E}(X^*w')}_{=0}\beta - \underbrace{\mathbb{E}(ww')}_{=\sigma_w^2 I}\beta \tag{243}$$

$$\therefore \quad \mathbb{E}(Xu^*) = cov(X, u^*) = cov(X + w, u - w\beta) = -\beta\sigma_w^2 I \tag{244}$$

Esto no satisface el supuesto de no correlación entre los regresores y el término de error. Entonces $\hat{\beta}$ será sesgado e inconsistente, de esto salen dos puntos importantes:

[20] También se puede comentar de que la asimetría es el 3er momento centrado y estandarizado de la serie.

- Asumir que la medida de los errores son no sistemático es ingenuo.
- Si ellos son sistemáticos, $\hat{\beta}$ será sesgado e inconsitente incluso en el primer caso.

3.7.1 Variable omitidas

Supondremos de que el modelo correcto es el siguiente:

$$Y = X_1\beta_1 + X_2\beta_2 + u \tag{245}$$

Y el modelo estimado es:

$$Y = X_1\beta_1 + u \tag{246}$$

En este caso por OLS, tendremos de que:

$$\hat{\beta_1} = (X_1'X_1)^{-1}X_1'Y = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'u \tag{247}$$

En este caso el estimador será insesgado en la mayoría de los casos, ya que:

$$\mathbb{E}(\hat{\beta_1}) = \beta_1 + \underbrace{(X_1'X_1)^{-1}X_1'X_2}_{Z}\beta_2 \tag{248}$$

Cada columna de Z será la derivada del regresor de X_1 y X_2, además se aproxima a las covarianzas, que sería como la pendiente de X_1 y X_2. La única forma de que $Z = 0$, es que X_1 y X_2 sean ortogonales o que $\beta_2 = 0$. La dirección del sesgo es difícil de evaluar en el caso generar, si suponemos de que ambos son escalares tendremos de que:

$$\mathbb{E}(\hat{\beta_1}) = \beta_1 + \frac{cov(X_1, X_2)}{V(X_1)} \cdot \beta_2 \tag{249}$$

Si el signo de $Cov(X_1, X_2)\beta_2 > 0$, $\mathbb{E}(\beta_1) > \beta_1$ y el estimador estará sobreestimando el efecto de X_1 en Y. En el caso de la varianza, nosotros estimaremos:

$$V(\hat{\beta_1}|X_1) = \sigma^2(X_1'X_1)^{-1} \tag{250}$$

Si denotamos al estimador correcto $\hat{\beta_1^*}$, tendremos que la varianza correcta es:

$$V(\hat{\beta_1^*}|X_1) = \sigma^2(X_1'M_2X_1)^{-1} \tag{251}$$

Para comparar ambas expresiones, calcularemos la inversa:

$$[V(\hat{\beta_1}|X_1)]^{-1} - [V(\hat{\beta_1^*}|X_1)]^{-1} = \sigma^{-2}(X_1'X_1) - \sigma^{-2}(X_1'M_2X_1) = \sigma^{-2}(X_1'X_2(X_2'X_2)^{-1}X_2'X_1) > 0 \tag{252}$$

Podemos concluir que aunque $\hat{\beta_1}$ es insesgado e inconsistente, tendrá menor varianza que $\hat{\beta_1^*}$. Además, tendremos de que si σ^2 no es conocida la tendremos que estimar, para hacerlo seguiremos pensando que el modelo estimado es correcto, tendremos de que:

$$\tilde{\sigma}^2 = \frac{\hat{u}'\hat{u}}{T - k_1} \tag{253}$$

Pero $\hat{u} = M_1Y = M_1(X_1\beta_1 + X_2\beta_2 + u) = M_1X_2\beta_2 + M_1u$, por lo tanto:

$$\mathbb{E}(\hat{u}'\hat{u}) = \mathbb{E}\left[(M_1X_2\beta_2 + M_1u)'(M_1X_2\beta_2 + M_1u)\right] = \beta_2'X_2'M_1X_2\beta_2 + \sigma^2(T - k_1) \tag{254}$$

El primer término es la contrapartida de la población al aumento del SSR debido a la omisión de X_2 de la regresión. Como el término $\tilde{\sigma}^2$ es positivo , estará sesgado hacia arriba (la verdadera varianza es menor). Desafortunadamente, para tener esto en cuenta deberíamos requerir conocer β_2. En conclusión:

- Si nosotros omitimos variables relevantes, $\hat{\beta_1}$ y $\tilde{\sigma}^2$ estarán sesgadas.
- Incluso cuando $\hat{\beta_1}$ sea más precisa que $\hat{\beta_1^*}$, $\tilde{\sigma}^2$ no será consistentemente estimada.
- Solo en el caso de que X_1 y X_2 son ortogonales, $\hat{\beta_1}$ será insesgada.

3.7.2 Variable irrelevante incluida

Supongamos ahora, de que el modelo real es de la forma:

$$Y = X_1\beta_1 + u \tag{255}$$

Pero nosotros estimamos:

$$Y = X_1\beta_1 + X_2\beta_2 + u \tag{256}$$

En este caso, nosotros estimamos:

$$\hat{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2Y = \beta_1 + (X_1'M_2X_1)^{-1}X_1'M_2u \tag{257}$$

La esperanza será de la forma:

$$\mathbb{E}(\hat{\beta}_1) = \beta_1$$

Por lo tanto:

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}\begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{bmatrix} = \begin{bmatrix}\beta_1 \\ 0\end{bmatrix} \tag{258}$$

Por la misma razón uno puede probar de que:

$$\mathbb{E}(\tilde{\sigma}^2) = \mathbb{E}\left(\frac{\hat{u}'\hat{u}}{T - k_1 - k_2}\right) = \sigma^2 \tag{259}$$

Entonces, ¿cuál es el problema del sobreajuste? La forma, el costo: la reducción en la precisión la cual recae en:

$$\hat{\beta}_1 = \beta_1 + (X_1'M_2X_1)^{-1}X_1M_2u \tag{260}$$

La varianza será de la forma:

$$V(\hat{\beta}_1|X)) = \sigma^2(X_1'M_2X_1)^{-1} \tag{261}$$

Esta varianza es mayor que si el modelo fuera estimado correctamente, ya que en este último caso:

$$V(\hat{\beta}_1^*|X_1) = \sigma^2(X_1'X_1)^{-1} \tag{262}$$

De esto se concluye lo siguiente:

- El estimador es asintomáticamente eficiente si X_1 y X_2 son ortogonales.
- En caso de que X_1 y X_2 estén áltamente correlacionados la inflación de la varianza sería alta.

3.8 Multicolinealidad

La multicolinealidad aparece cuando dos o más regresores están altamente correlacionados, lo que implica un difultad al estimar la precisión de los efectos individuales de cada uno.

3.8.1 Perfecta colinealidad

Esto es un fenómeno el cual no es un problema en si, sino más bien un descuido debido a la mala especificación. Sucede cuando las columnas de X son linealmente dependientes entre si, esto pasa cuando son de rango incompleto (es decir $rank(X'X) < k$), por lo tanto $\hat{\beta}$ no está definido. Esto sugiere de que hay una variable que está demás.
Este error es sencillo de identificar, dado de que $(X'X)$ no será invertible y por ende no existirá $(X'X)^{-1}$

3.8.2 Multicolinealidad débil

A diferencia de la colinealidad perfecta, la colinealidad débil presenta un problema estadístico , el problema no es de identificación sino que de precisión, ya que ante una alta correlación entre regresores habrá una baja precisión de la estimación estimada. Los síntomas del problema son:

- Pequeños cambios en los datos producen grandes cambios en las estimaciones.
- Los estadísticos t no son significativos, pero el R^2 es alto.
- Los coeficientes tienen signos incorrectos o las magnitudes son inverosímiles.

El problema aparece cuando $X'X$ es "débil singular" y las columnas de las X están cerca de ser linealmente dependientes, esto hace de que la confiabilidad de los cálculos son reducidos, esto debido al "floating-point" que dificulta el cálculo.
Como el problema es con $(X'X)^{-1}$, lo veremos al detalle. El elemento $j-th$ de la diagonal de $(X'X)^{-1}$ es (tomaremos $j=1$ por conveniencia):

$$(x_1'M_2x_1)^{-1} = (\underbrace{x_1'x_1}_{\text{escalar}} - x_1'X_2(X_2'X_2)^{-1}X_2'x_1)^{-1}$$

Dado de que $(x_1'M_2x_1)^{-1}$ es un escalar:

$$= (x_1'x_1(1 - \frac{x_1'\overbrace{X_2(X_2'X_2)^{-1}X_2'}^{P_{x_2}}x_1}{x_1'x_1}))^{-1} = (x_1'x_1(1 - \frac{x_1'P_{x_2}x_1}{x_1'x_1}))^{-1} \tag{263}$$

Es sencillo observar de que $\frac{x_1'X_2(X_2'X_2)^{-1}X_2'x_1}{x_1'x_1}$ se aproxima a la suma de cuadrado de los x_1, donde $(X_2'X_2)^{-1}X_2'x_1$ es la regresión de x_1 respecto a X_2, por lo tanto $X_2(X_2'X_2)^{-1}X_2'x_1$ es el valor predicho de los $x_1 \longrightarrow \hat{x}_1$, por lo tanto[21]:

$$(x_1'M_2x_1)^{-1} = \frac{1}{x_1'x_1(1-R^2_{1,NC})} \tag{264}$$

En donde X_2 es la matriz $T \times (k-1)$ de X que excluye a x_1 y R_1^2 es el R^2(no centrado) de la regresión de x_1 en el otro regresor.

Otra manera de derivar esto, es teniendo la siguiente regresión:

$$Y = \beta_1 x_1 + \beta_2 X_2 + \varepsilon \tag{265}$$

Donde x_1 es un vector y X_2 es una matriz, usando herramientas de regresión particionada, tendremos de que:

$$Var(\hat{\beta}_1|x_1, X_2) = Var((x_1'M_2x_1)^{-1}(x_1'M_2Y)|x_1, X_2) = \sigma^2(x_1'M_2x_1)^{-1} \tag{266}$$

Como ya derivamos el lado derecho, tendremos que la varianza es:

$$Var(\hat{\beta}_1|x_1, X_2) = \frac{\sigma^2}{x_1'x_1(1-R^2_{1,NC})} \tag{267}$$

En el caso de que x_1 y X_2 son ortogonales, $R^2_{1,NC} = 0$ y la varianza es la típica $Var(\hat{\beta}_1|x_1, X_2) = \sigma^2(x_1'x_1)^{-1}$. Si tenemos que una serie de regresores están áltamente correlacionados a x_1, entonces $R^2_{1,NC}$ tenderá a 1 y

[21] Esto se desprende de que $R^2_{NC} = \frac{\hat{Y}'\hat{Y}}{Y'Y} = \frac{Y'P_xY}{Y'Y}$ como se puede observar, es distinto al R^2, en este caso es como si se regresionara X_1 en X_2.

$Var(\hat{\beta}_1|x_1, X_2) \longrightarrow \infty$. De esto último se deriva el test VIF (variance inflation factor), el cual se construye de la forma:

$$VIF = \frac{1}{1 - R_j^2} \tag{268}$$

La regla general es procuparse cuando $R^2 < R_j^2 \ \forall \ j$, este "test" no es uno propiamente tal, ya que no lo comparamos con una distribución, sino que solo lo usamos como una regla "'al ojo" y decimos que hay problemas de colinealidad cuando $VIF > 10$.
Una medida alternativa es el test Besley, el cual se basa e condicionar un número (γ), el cual se construye con los valores propios de una matriz $B = S(X'X)S$, en donde S es una matriz diagonal de la forma:

$$S = \begin{bmatrix} \frac{1}{\sqrt{x_1'x_1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{x_2'x_2}} & 0 & \vdots \\ 0 & \cdots & 0 & \frac{1}{\sqrt{x_k'x_k}} \end{bmatrix} \tag{269}$$

Por lo tanto, la matriz B es de la forma (Para simplificar usaremos dos regresores):

$$\begin{aligned} B &= \begin{bmatrix} \frac{1}{\sqrt{X_1'X_1}} & 0 \\ 0 & \frac{1}{\sqrt{X_2'X_2}} \end{bmatrix} \begin{bmatrix} X_1' \\ X_2' \end{bmatrix} \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{X_1'X_1}} & 0 \\ 0 & \frac{1}{\sqrt{X_2'X_2}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{X_1}{\sqrt{X_1'X_1}} \\ \frac{X_2}{\sqrt{X_2'X_2}} \end{bmatrix} \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{X_1'X_1}} & 0 \\ 0 & \frac{1}{\sqrt{X_2'X_2}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{X_1'X_1}{\sqrt{X_1'X_1}} & \frac{X_1'X_2}{\sqrt{X_1'X_1}} \\ \frac{X_2'X_1}{\sqrt{X_2'X_2}} & \frac{X_2'X_2}{\sqrt{X_2'X_2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{X_1'X_1}} & 0 \\ 0 & \frac{1}{\sqrt{X_2'X_2}} \end{bmatrix} \end{aligned}$$

$$B = \begin{bmatrix} \frac{X_1'X_1}{X_1'X_1} & \frac{X_1'X_2}{\sqrt{X_1'X_1}\sqrt{X_2'X_2}} \\ \frac{X_2'X_1}{\sqrt{X_2'X_2}\sqrt{X_1'X_1}} & \frac{X_2'X_2}{X_2'X_2} \end{bmatrix} = \begin{bmatrix} 1 & \frac{X_1'X_2}{\sqrt{X_1'X_1}\sqrt{X_2'X_2}} \\ \frac{X_2'X_1}{\sqrt{X_2'X_2}\sqrt{X_1'X_1}} & 1 \end{bmatrix} \tag{270}$$

Es sencillo ver, de que cuando no hay colinealidad X_1 es ortogonal a X_2 y por ende $X_1'X_2 = 0$, de esta forma la matriz de Besley es identidad:

$$B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{271}$$

En este caso los eigenvalues, son iguales a 1, ya que:

$$det(B - \lambda I) = 0 \longrightarrow det \begin{bmatrix} 1 - \lambda & 0 \\ 0 & 1 - \lambda \end{bmatrix} = (1 - \lambda)^2 = 0$$

$$\therefore \quad \lambda = 1$$

En el caso de perfecta colinealidad, el determinante será igual a 0, ya que las cercanías de la diagonal serán iguales a 1, el test Besley establece un número (como se dijo al principio), el cual depende de los eigenvalues de la siguiente forma:

$$\gamma = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \tag{272}$$

Como hemos comentado con anterioridad, cuando no hay colinealidad los valores propios son iguales a 1 y por ende $\gamma = 1$, en el caso de que hay colinealidad perfecta los valores propios son iguales a 0 y por ende $\gamma \longrightarrow \infty$, Besley sugiere de que cuando $\gamma > 20$ indica de que hay potenciales problemas, existen algunas maneras de lidiar con "el problema":

- Reducir las dimensiones de X (dropear variables), el problema de hacer esto es que como la variable es relevante, $\hat{\beta}$ estará sesgada (como vimos con anterioridad). Esta práctica hace explícito el *trade-off* entre reducción de varianza y sesgo.
- Utilizar componente principales.
- utilizar Ridge regression.

En resumen no hay palabra más mal usada que "Problema de multicolinealidad", puesto de que si las variables están altamente colineales es un hecho de la causa, la contemplación acerca de la aparente malevolencia de la naturaleza no es contructivo (además de no ser correcto), las curas ad-hoc para una mala muestra pueden ser desastrozamente inapropiadas (intentando ganar eficiencia, perdemos consistencia). Por lo tanto, es mejor aceptar el hecho de que los datos no experimentales en ocasiones no son muy informativo de los parámentros de interés.

3.8.3 Componente principal

Tome el modelo

$$Y = X\beta + u \tag{273}$$

Considere la transformación[22]:

$$Y = \underbrace{XP}_{=Z}\underbrace{P'\beta}_{=\theta} + u = Z\theta + u \tag{274}$$

Donde:

$$P_{k\times k} = \begin{bmatrix} p_1 & \cdots & p_k \end{bmatrix} \tag{275}$$

p_j es el eigenvector $j-th$ ortogonal (vector característico) de $X'X$. Estos eigenvectors son ordenados por el orden de la magnitud del correspondiente eigenvalue, por lo tanto:

$$Z_{T\times k} = \begin{bmatrix} z_1 & \cdots & z_k \end{bmatrix} \tag{276}$$

Esta es la matriz del componente principal, es decir $z_j = Xp_j$, la cual es llamada como el componente principal $j-th$, donde $z_j'z_j = \lambda_j$[23].
El estimador del principal componente de β es obtenido mediante la eliminación de uno o más z_j, aplicando OLS para reducir el modelo y transformando el estimador obtenido al espacio del parámetro original. Esto es, particionando $X\begin{bmatrix} P_1 & P_2 \end{bmatrix} = \begin{bmatrix} Z_1 & Z_2 \end{bmatrix}$, tendremos:

$$Y = X_1P_1\theta_1 + XP_2\theta_2 + u = Z_1\theta_1 + Z_2\theta_2 + u \tag{277}$$

Si omitimos Z_2 del modelo obtenemos, vía OLS:

$$\hat{\theta}_1 = (Z_1'Z_1)^{-1}Z_1'Y$$

Como Z_1 y Z_2 son ortogonales $\hat{\theta}_1$ está insesgado. Por lo tanto, $V(\hat{\theta}_1) = \sigma^2(Z_1'Z_1)^{-1}$. Este estimador tiene las propiedades deseables para θ_1, pero no para los parámetros reales de interés (β). Ahora discutiremos una transformación de θ_1 devuelta en β que se propone usualmente, note de que $\beta = P\theta = P_1\theta_1 + P_2\theta_2$, como omitimos Z_2 nosotros explícitamente asumimos de que θ_2 es igual a 0, en estos casos $\hat{\beta}^* = P_1\hat{\theta}_1$, el cual será el estimador del componente principal de β. Como vimos anteriormente (en variables omitidas), es sencillo notar de que $V(\hat{\beta}^*) < V(\hat{\beta})$, sin embargo este estimador estará sesgado a menos de que $P_2\theta_2 = 0$.
Hasta ahora, nosotros hemos guardado silencio acerca de cómo escoger Z_2. Dos enfoques han sido sugeridos:

- Incluir en Z_2 el componente con el valor propio más pequeño. Esto equivale a decir que colinealidad débil es equivalente a colinealidad perfecta, lo cual no es muy recomendable.
- El test para $P_2'\theta_2 = 0$ (lo cual no es trivial de probar).

[22] Recuerde que la matriz P satisface $P'P = PP' = I$.
[23] λ_j denota el mayor eigenvalue jth de $X'X$

3.8.4 Ridge Regression

Teniendo $\Lambda = P'X'XP = diag(\lambda_1, \ldots, \lambda_k)$, como la matriz diagonal de valores propios de $X'X$ (como lo vimos antes, es la matriz de eigenvectors de $X'X$). El estimador de la regresión Rigde generalizada (GRR) está definida por:

$$\tilde{\theta} = (\Lambda + W)^{-1} Z'Y = (\Lambda + W)^{-1} \Lambda \hat{\theta} \tag{278}$$

Donde:

$$W = diag(w_1, \ldots, w_k), \quad w_i > 0, \tag{279}$$

y :

$$\hat{\theta} = \Lambda^{-1} Z'Y \tag{280}$$

Es la estimación OLS de θ. Recuerde que $\theta = P'\beta$, el estimador GRR de β es:

$$\tilde{\beta} = P\tilde{\theta} \tag{281}$$

El estimador GRR depende de la elección de W. Se puede demostrar, de que los valores de w_i que minimizan el error cuadrático medio (MSE) de $\tilde{\beta}$ están dados por:

$$w_i = \frac{\sigma^2}{\theta_1^2} \tag{282}$$

Donde θ_i es el elemento i-th de θ, se puede obtener un estimador operacional sustituyendo σ^2 y θ_i, vía OLS de la forma:

$$\hat{w}_i = \frac{\tilde{\sigma}^2}{\hat{\theta}_i^2} \tag{283}$$

Una versión simple del estimador, es llamado el estimador ordinario de la regresión Ridge (ORR), el cual es obtenido al establecer $W = wI$:

$$\tilde{\beta}_{ORR} = (X'X + wI)^{-1} X'Y \tag{284}$$

Aunque no se puede encontrar un valor óptimo explícito para w, se han propuesto varias opciones estocásticas. Entre los más populares tenemos:

$$\hat{w}^I = \frac{k\tilde{\sigma}^2}{\hat{\beta}'\hat{\beta}} \quad y \quad \hat{w}^{II} = \frac{k\tilde{\sigma}^2}{\hat{\beta}'X'X\hat{\beta}} \tag{285}$$

Aunque los estimadores de la regresión Ridge, pueden tener un MSE menor que OLS, es importante mencionar varios incovenientes:

- No hay consenso con respecto a la elección de parámetros de contracción.
- El parámetro de contracción no tiene distribución estándar.
- Debido a esto, la distribución del estimador de la regresión Ridge de β además no será estándar, en cuyo caso la inferencia no puede realizarse con las pruebas iniciales (especialmente en muestras pequeñas).

3.8.5 Volviendo a la discusión anterior

Para ejemplificar, el por qué argumentamos de que el "problema de multicolinealidad", no es un problema ad-hoc, considere lo siguiente:

$$y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + u_t \tag{286}$$

Una regresión de x_2 en x_1 llegamos a $x_{2,t} = \hat{\theta} x_{1,t} + \hat{v}_t$ donde $\hat{v}$ (por construcción, es ortogonal a x_1), sustituyendo este auxiliar en la regresión original, obtenemos el siguiente modelo:

$$y_t = \beta_1 x_{1,t} + \beta_2(\hat{\theta} x_{1,t} + \hat{v}_t) + u_t = (\beta_1 + \beta_2\hat{\theta}) x_{1,t} + \beta_2 \hat{v}_t + u_t = \delta_1 z_{1,t} + \delta_2 z_{1,t} + u_t \tag{287}$$

Donde $(\beta_1 + \beta_2\hat{\theta}) = \delta_1$, $\delta_2 = \beta_2$, $z_1 = x_1$ y $z_2 = x_2 + \hat{\theta}x_1$. El investigador quien usa las variables x_1 x_2 con los parámetros β_1 y β_2, reportara que β_2 está estimado incorrecto por el problema de multicolinealidad. Pero otro investigador el cual tropezará al estimar el modelo con las variables z_1 y z_2 con los parámetros θ_1 y θ_2 reportaría que este no es un problema de multicolinealidad, debido a que z_1 y z_2 son ortogonales (recuerde de que x_1 y $\hat{v}$ son ortogonales).No obstante, este investigador reportaría de que $\gamma_2 (= \beta_2)$ está estimado de manera inexacta, no por colinealidad, sino porque z_2 no está variando adecuadamente.