# Tercer ejercicio

El objetivo de este ejercicio es indagar sobre las distintas estructuras de errores estándar que comúnmente se emplean al estimar un modelo de regresión lineal. En particular, queremos evaluar cómo distintas metodologías para la estimación de los errores estándar se comportan en diversos contextos que pueden hacer que uno u otro enfoque resulte más apropiado. El análisis se centrará en estimar y efectuar inferencia estadística sobre el siguiente modelo:

$$Y_i = \beta * X_i + \varepsilon_i, \quad \beta = 10 \tag{0.1}$$

Donde inicialmente los $X_i$ son independientes e idénticamente distribuidos con distribución normal estándar, mientras que en los siguientes puntos variaremos la estructura de la distribución de los términos de error $\varepsilon_i$.

1. Inicialmente, considere un escenario homocedástico en el que $\varepsilon_i \sim \mathcal{N}(0,1)$ iid. En este contexto, estudiaremos los efectos de suponer erróneamente un modelo heterocedástico (pero sin autocorrelación).

   a) En primer lugar, consideraremos dos posibles estimadores de la varianza del vector de $\hat{\beta}$, que viene dada por

   $$var(\hat{\beta}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$$

   Donde $\Sigma$ es la matriz de varianzas y covarianzas del vector de $\varepsilon_i$'s y $X$ es la matriz de variables explicativas.

   Bajo errores homocedásticos, se supone $\Sigma = \sigma^2 I$, de manera que solo es necesario estimar $\sigma^2$, lo cual se hace por medio de la fórmula

   $$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-k}$$

   donde $k$ es el número de parámetros y $\hat{\varepsilon}_i$ son los residuales de la regresión.

   En cambio, bajo errores heterocedásticos, se hace necesario estimar directamente la matriz de varianzas y covarianzas $\Sigma$. Una alternativa usual son los errores tipo Huber-White o *hc1* (*heteroskedasticity consistent 1*, los cuales son la opción predeterminada de Stata al agregar la opción "*, robust*"). En particular este estimador toma la forma:

   $$\hat{var}(\hat{\beta})_{hc1} = (X'X)^{-1}X'\text{diag}(\hat{\epsilon}_1^2, \cdots, \hat{\epsilon}_n^2)X(X'X)^{-1}$$

   donde el estimador de la matriz $\Sigma$ es la matriz diagonal compuesta por los residuales del modelo al cuadrado ($\hat{\Sigma}_{hc1} = \text{diag}(\hat{\epsilon}_1^2, \cdots, \hat{\epsilon}_n^2)$). En otras palabras, el estimador puntual del elemento $\Sigma_{ii}$ es $\hat{\varepsilon}_i^2$.

a) Simulen 1000 muestras $\{X_i, Y_i\}_{n=1}^{1000}$ acorde al modelo (1). Para cada una, estimen el modelo (1) y calculen el intervalo de confianza del 95 % para $\hat{\beta}$ empleando errores estándar usuales y *hc1*. Para cada caso, calcule el porcentaje de las veces en que en dicho intervalo se encuentra el parámetro verdadero. Presenten e interpreten sus resultados.

Tabla 1: 1000 observaciones

| VARIABLES | (1) Errores estandar | (2) hc1 |
|---|---|---|
| % de éxito del parámetro en intervalo | 94.7 % | 94.3 % |
| Número de observaciones | 1000 | 1000 |

Los intervalos de confianza y los errores estándar son muy parecidos, por tanto, se podría decir que no se tienen unos cambios muy importantes entre los errores estandar y los robustos a heterocedasticidad cuando se tiene una muestra homocedástica. Pues el aumento en la precisión del parámetro en el intervalo es muy pequeño.

b) Los errores $hc1$ resultan problemáticos cuando se cuenta con muestras pequeñas por dos razones principales. Por una parte, aunque el estimador obtenido de $\Sigma$ es consistente, su convergencia es relativamente lenta. Por otra parte, se ha probado que es especialmente sensible a la presencia de datos atípicos. Para visualizar estos problemas, repitan el procedimiento del inciso anterior, esta vez con muestras de tamaño 10. ¿Cómo se comparan sus resultados?

Tabla 2: 10 observaciones

| VARIABLES | (1) Errores estandar | (2) hc1 |
|---|---|---|
| % de éxito del parámetro en intervalo | 100 % | 90 % |
| Número de observaciones | 10 | 10 |

Se presentan cambios mayores en el porcentaje de éxito del parámetro, en este caso se puede ver que al usar errores robustos se tienen menos casos de éxito respecto a los errores estándar normales. Esto puede dar cuenta de las falencias de los errores hc1 con algunos datos atípicos que toman más relevancia cuando se tiene una muestra pequeña.

c) Una solución al problema de muestra pequeña al que se enfrentan los errores *hc1* fue propuesta por Davidson y MacKinnon (1993), quienes sugieren usar un estimador alternativo que contiene un factor de corrección para muestras finitas[1] $\hat{\Sigma}_{hc3} = \text{diag}(\hat{\epsilon}_1^2/(1-h_{11})^2, \cdots, \hat{\epsilon}_n^2/(1-h_{nn})^2)$, en donde $h_{ii}$ se refiere al i-ésimo elemento de la diagonal de la matriz $X(X'X)^{-1}X$. La intuición detrás de esta corrección es que cuantifica, en cierta medida, qué tan influyente es una observación a la hora de calcular $\hat{\beta}$. De esta manera, este estimador se hace más resistente a datos atípicos[2]. Repita el procedimiento del inciso $b$, pero esta vez emplee errores convencionales y $hc3$. Reporte sus resultados y compare.

Tabla 3: 10 observaciones

| VARIABLES | (1) Errores estandar | (2) hc3 |
|---|---|---|
| % de éxito del parámetro en intervalo | 94.7 % | 93.8 % |
| Número de observaciones | 1000 | 1000 |

En este caso los errores estándar robustos hc3 presentan una mayor variación en el porcentaje de éxito respecto al de los errores hc1, en este caso se tiene una diferencia de casi un punto porcentual, mientras que con los errores hc1 no se veían cambios significativos.

2. Suponga ahora que $\varepsilon_i \sim \mathcal{N}(0, X_i^2)$ independientes.

a) Simulen 1000 muestras $\{X_i, Y_i\}_{n=1}^{1000}$ acorde al modelo (1). Para cada una, estimen el modelo (1) y calculen el intervalo de confianza $\hat{\beta}$ empleando errores estándar usuales, *hc1* y *hc3*. Para cada caso, calcule el porcentaje de las veces en que en dicho intervalo se encuentra el parámetro verdadero. Presenten e interpreten sus resultados. Comparen con lo obtenido en el punto 1.

Tabla 4: 1000 observaciones

| VARIABLES | (1) Errores estandar | (2) hc1 | (3) hc3 |
|---|---|---|---|
| % de éxito del parámetro en intervalo | 62.4 % | 94.4 % | 94.7 % |
| Número de observaciones | 1000 | 1000 | 1000 |

En el caso de esta distribución, se tienen mejoras muy considerables con errores estándar robustos (hc1) y los errores estándar robustos de Davidson y MacKinnon (hc3). El porcentaje de éxito cambia en casi 32 puntos porcentuales respecto a los errores estandar normales. Por otro lado, se puede ver que las diferencias entre hc1 y hc3 no son tan grandes, lo que puede dar una visión de que los cambios fundamentales entre estos dos métodos ocurren con muestras pequeñas.

[1] Puede revisar este link o este para más información.
[2] Este procedimiento es, de hecho, el default en R al usar errores robustos.

b) Repitan el anterior procedimiento con muestras de tamaño 10. Presenten e interpreten sus resultados. Comparen con lo obtenido en el punto 1.

Tabla 5: 10 observaciones

| VARIABLES | (1) Errores estandar | (2) hc1 | (3) hc3 |
|---|---|---|---|
| % de éxito del parámetro en intervalo | 40 % | 90 % | 100 % |
| Número de observaciones | 10 | 10 | 10 |

En este caso resaltan las diferencias entre los errores estandar tradicionales son más grandes respecto al caso de 1000 muestras, y además, se evidencia una mayor diferencia en el porcentaje de éxito entre los errores hc1 y los hc3, siendo estos últimos los que mejores resultados se obtuvieron.

3. Otro problema común es la autocorrelación o correlación serial, la cual tiene que ver con la existencia de grupos de unidades u observaciones dentro de la muestra cuyos términos de error se encuentran correlacionados. Este panorama rompe con el supuesto clásico de independencia entre los términos de error y plantea nuevos problemas que exploraremos a continuación.

   a) Simulen 1000 muestras $\{X_i, Y_i\}_{n=1}^{1000}$ acorde a los siguientes pasos:

      - Simule las primeras 100 observaciones de la muestra siguiendo el modelo 1 tomando $\varepsilon_i \sim \mathcal{N}(0,1)$ iid.
      - Por medio del comando **expand**, produzca 10 copias de los datos (de esta forma construya una base con 1000 observaciones en total)

      Explique por qué este procedimiento induce un problema de autocorrelación. Para cada muestra, estimen el modelo (1) y calculen el intervalo de confianza $\hat{\beta}$ empleando errores estándar usuales, *hc1*, *hc3* y cluster [3]. Los errores cluster permiten que las covarianzas de errores de observaciones pertenecientes a un mismo cluster sean estimadas individualmente (i.e. da libertad sobre el valor de estas covarianzas), mientras que las demás covarianzas son restringidas a 0. En cada caso, calcule el porcentaje de las veces en que en dicho intervalo se encuentra el parámetro verdadero. Presenten e interpreten sus resultados.

Tabla 6: 1000 observaciones

| VARIABLES | (1) Errores estandar | (2) hc1 | (3) hc3 | (4) cluster |
|---|---|---|---|---|
| % de éxito del parámetro en intervalo | 46 % | 50 % | 43 % | 93 % |
| Número de observaciones | 1000 | 1000 | 1000 | 1000 |

Esto genera un problema de autocorrelación porque al repetir 10 veces los valores generados inicialmente voy a tener muestra que depende de una observación anterior. Al ser observaciones iguales el error puede estar correlacionado con una de las creadas repetidamente pues se repiten secuencialmente en mis datos, (primero 10 veces el primer elemento inicial, luego 10 veces el segundo y así sucesivamente) lo que genera que los errores de las observaciones adyacentes esten correlacionados, estas son las que debemos seleccionar para hacer los clusters. Como se puede ver en los resultados de la columna 4, al hacer los clusters se obtienen mejoras sustanciales en el porcentaje de éxito del parámetro.

   b) ¿Es el modelo heterocedástico sin autocorrelación un caso específico de errores con estructura cluster?

      Si, pues al tratarse de un modelo heterocedástico la varianza de los errores es distinto para cada x, y en la estructura cluster de los errores si la varianza del estimador agrupado es menor que la del estimador no agrupado, significa que las sumas agrupadas de ei*xi tienen menos variabilidad que los ei*xi individuales. Es decir, cuando se suman los ei*xi dentro de un conglomerado, parte de la variación se anula y la variación total es menor.

   c) Un problema con los errores clusters es que el número de entradas de la matriz varcov de los errores que hay que estimar puede ser muy alto si el número de clusters es relativamente bajo.

[3] *Pista:* piense cuáles son las observaciones que se correlacionan para escoger el nivel de cluster

Como consecuencia, con pocos clusters se hace necesario estimar un mayor número de parámetros simultáneamente, lo que puede afectar la calidad de dicha estimación. Para ver esto, simulen 1000 muestras $\{X_i, Y_i\}_{n=1}^{1000}$ para $j = 10, 20, 50, 100$ según el siguiente procedimiento:

- Simulen las primeras $j$ observaciones.
- Dupliquen las base de datos tantas veces sea necesario para producir 1000 observaciones en total.
- Para cada muestra, estimen el modelo (1) y calculen el intervalo de confianza $\hat{\beta}$ empleando errores estándar cluster.

Para cada $j$, calcule el porcentaje de las veces en que en dicho intervalo se encuentra el parámetro verdadero. Gráfique el $j$ empleado contra el porcentaje de aciertos respectivo e interprete.

4. Una alternativa al problema de seleccionar la estructura de los errores para estimar su matriz de varianza y covarianza es a través de métodos no paramétricos. En particular, la metodología de bootstrap nos permite recuperar una estimación del error estándar incluso en contextos en los que no es claro cuál de todos los estimadores previamente vistos es más adecuado.

   a) Simulen 1000 muestras $\{X_i, Y_i\}_{n=1}^{1000}$ acorde al modelo (1) cuando $\varepsilon_i \sim \mathcal{N}(0, 1)$ iid. Para cada una, estimen el modelo (1) y calculen el intervalo de confianza $\hat{\beta}$ del 95 % empleando errores estándar bootstrap con 100 repeticiones. Calculen el porcentaje de las veces en que en dicho intervalo se encuentra el parámetro verdadero.

Tabla 7: 1000 observaciones

| VARIABLES | (1)<br>Errores estandar |
|---|---|
| % de éxito del parámetro en intervalo | 95.2 % |
| Número de observaciones | 1000 |

   b) Simulen 1000 muestras $\{X_i, Y_i\}_{n=1}^{1000}$ acorde al modelo (1) cuando $\varepsilon_i \sim \mathcal{N}(0, X_i^2)$ independientes. Para cada una, estimen el modelo (1) y calculen el intervalo de confianza $\hat{\beta}$ del 95 % empleando errores estándar bootstrap con 100 repeticiones. Calculen el porcentaje de las veces en que en dicho intervalo se encuentra el parámetro verdadero.

Tabla 8: 1000 observaciones

| VARIABLES | (1)<br>Errores estandar |
|---|---|
| % de éxito del parámetro en intervalo | 93.7 % |
| Número de observaciones | 1000 |

   c) Simulen 1000 muestras $\{X_i, Y_i\}_{n=1}^{1000}$ acorde a los siguientes pasos:

      - Simule las primeras 100 observaciones de la muestra siguiendo el modelo 1 tomando $\varepsilon_i \sim \mathcal{N}(0, 1)$ iid.
      - Por medio del comando **expand**, produzca 10 copias de los datos (de esta forma construya una base con 1000 observaciones en total)

      Para cada muestra, estimen el modelo (1) y calculen el intervalo de confianza $\hat{\beta}$ del 95 % empleando errores estándar bootstrap con 100 repeticiones. Calculen el porcentaje de las veces en que en dicho intervalo se encuentra el parámetro verdadero.

Tabla 9: 1000 observaciones

| VARIABLES | (1)<br>Errores estandar |
|---|---|
| % de éxito del parámetro en intervalo | 39 % |
| Número de observaciones | 1000 |

d) Comparen sus resultados con lo obtenido en los anteriores incisos.

Para comenzar podemos ver que los resultados del 4.a son similares a los del 1.a en el que se había obtenido un porcentaje de acierto del 94.7 %, mientras que utilizando los errores bootstrap se tiene un acierto del 95.2 %. Por otro lado, en el caso de la distribución con desviación estándar $x^2$, obtuvo mejores resultados que utilizando los errores estandar clásicos y resultados similares de solo un punto porcentual de diferencia respecto a los errores robustos hc1 y hc3. Finalmente, donde más se tuvieron diferencias y problemas en el porcentaje de éxito del parámetro en el intervalo fue en el caso donde se crean 100 observaciones y luego se reproducen copias exactas debido a que este solo tuvo un 39 % de éxito, teniendo el nivel más bajo respecto a los otros cuatro tipos de errores. En este caso, el que mejores resultados obtuve fue el de cluster, mientras que los otros, inclído el bootstrap tuvieron porcentajes de éxito muy bajos.