Serie Notas Técnicas
Superintendencia de Pensiones
Teatinos 317.
Santiago, Chile.

www.spensiones.cl

NOTA TÉCNICA N° 3

EFECTO DE REDONDEAR VARIABLES MONETARIAS EN BASES DE DATOS ADMINISTRATIVAS DE LA SUPERINTENDENCIA DE PENSIONES

Ximena Quintanilla

Mayo 2010

Las **Notas Técnicas** son una línea de publicaciones de la Superintendencia de Pensiones, que tienen por objeto divulgar artículos breves escritos por profesionales de esta institución.

Los trabajos aquí publicados tienen carácter preliminar y están disponibles para su discusión y comentarios. Los contenidos, análisis y conclusiones que puedan derivar de los documentos publicados son de exclusiva responsabilidad de su(s) autor(es) y no reflejan necesariamente la opinión de la Superintendencia de Pensiones.

Si requiere de mayor información o desea tomar contacto con quienes editan estos documentos, contacte a: documentosdetrabajo@spensiones.cl.

Si desea acceder a los títulos ya publicados y/o recibir las futuras publicaciones, por favor regístrese en nuestro sitio web: www.spensiones.cl.

Superintendencia de Pensiones
Teatinos 317
Santiago 8340382, Chile.
www.spensiones.cl

Efecto de Redondear Variables Monetarias en Bases de Datos Administrativas de la Superintendencia de Pensiones

Ximena Quintanilla*
Superintendencia de Pensiones

Mayo 2010

Resumen

La Superintendencia de Pensiones pone a disposición de la comunidad académica y diseñadores de políticas públicas muestras representativas de los afiliados al Sistema de Pensiones y al Seguro de Cesantía. Entre las medidas adoptadas por la Superintendencia para proteger la identidad de los individuos se encuentra el redondeo de cada variable monetaria al 0.25 % de su promedio. Este artículo analiza el efecto de esta medida en las estimaciones de medias y desviaciones estándar en el caso univariado y para un sencillo ejemplo de regresión multivariada.

1. Introducción

La Superintendencia de Pensiones recibe, regularmente, bases de datos desde las instituciones reguladas como las Administradoras de Fondos de Pensiones y la Administradora de Fondos de Cesantía. Estas bases de datos contienen información mensual desagregada a nivel individual de cotizaciones en el respectivo sistema, ingreso imponible, monto de los beneficios para aquellos que los reciben, monto acumulado en la cuenta individual, etc.

El carácter individual y la precisión de los datos administrativos implican que estas bases de datos son una fuente valiosa para realizar estudios y análisis empíricos sobre el sistema de pensiones, el seguro de cesantía y el mercado del trabajo. En particular, la Ley 20.328 que en el año 2009 reformó el Seguro de Cesantía, reconoce este hecho estableciendo que la Superintendencia deberá elaborar una muestra representativa de los afiliados al Seguro de Cesantía. Así, con el objetivo de aumentar el conocimiento y promover un debate informado, la Superintendencia de Pensiones pone la muestra de Afiliados al Seguro de Cesantía a disposición de la comunidad académica, diseñadores de políticas públicas y agentes relacionados.

*Analista Senior, División Estudios Superintendencia de Pensiones. xquintanilla@spensiones.cl. Se agradecen los comentarios de Gonzalo Reyes, Olga Fuentes y Nicolás Torrealba.

Dado que la bases de datos de la Superintendencia de Pensiones contienen información altamente sensible sobre los afiliados a los distintos sistemas, la Superintendencia ha tomado diversas medidas para minimizar la posibilidad de identificar a los individuos en la bases de datos. Esta nota técnica detalla dichas medidas y analiza el efecto que una de ellas, redondear las variables monetarias, puede tener en las estimaciones. El documento también propone y evalúa un método de corrección disponible en la literatura para datos redondeados.

La estructura de la nota es la siguiente: la próxima sección describe las razones legales para proteger los datos de la posibilidad de identificar afiliados. La sección 3 especifica las medidas adoptadas para proteger los datos, incluyendo la regla de redondeo. La sección 4 utiliza la base de datos de Historias Previsionales de Adminsitrativas, HPA, para proveer evidencia del efecto que la regla de redondeo puede tener en estadísticas y estimaciones. La sección 5 concluye.

2. Necesidad de proteger los datos

Por un lado, la Ley 19.628 permite a las instituciones públicas utilizar bases de datos para desarrollar sus tareas, que en el caso de la Superintendencia incluyen ."Efectuar los estudios técnicos y actuariales necesarios para el ejercicio de sus atribuciones"[1]. Además, la Ley 20.328 establece que la Superintendencia de Pensiones pondrá una muestra representativa de afiliados al Seguro de Cesantía a disposición de la comunidad académica, investigadores y personas afines para que realicen estudios y para que los diseñadores de políticas públicas tengas mejores herramientas para diseñar y evaluar políticas.

Por otro lado, la Ley 19.628 intenta proteger la identidad e integridad de aquellos cuya información está en las bases de datos. Específicamente, la Ley menciona que los individuos que trabajen utilizando bases que contengan datos personales, tanto en instituciones privadas como públicas, están obligados a mantener la más estricta confidencialidad sobre los datos. Además, los responsables de las bases donde se almacenen datos personales debe resguardar la información y serás responsables de cualquier daño derivado de ellas.

El enfoque que la Superintendencia ha adoptado para conciliar el deseo de permitir el uso de la base de datos para efectos de investigación con el mandato de proteger la información personal sensible es hacer pública la infromación adoptando varias medidas para evitar la identificación de los individuos.

3. Medidas para proteger los datos

Las medidas adoptadas para minimizar la probabilidad de identificar individuos en las muestras de afiliados son:

[1] Ley 20.255, Artículo No. 47.

1. Reemplazo del RUT del individuo por un número de registro único. Sólo la Superintendencia tendrá acceso a la llave que relaciona los dos identificadores
2. Reemplazo del RUT del empleador por un número de registro único. Sólo la Superintendencia tendrá acceso a la llave que relaciona los dos identificadores
3. Las fechas incluyen sólo mes y año, se omite el día
4. Se omite la dirección específica del individuos, solo se reporta la variable Comuna.
5. Cada variable monetaria es redondeada a 0.0025 de su promedio[2].

La regla de redondeo aplicada se basa en la metodología seguida por el Health and Retirement Study a datos vinculados con la registros administrativos de la Seguridad Social.

4. Efecto de redondear los datos

En esta sección se utiliza la base de datos de Historias Previsionales Administrativas, HPA, para analizar el efecto de redondear las variables monetarias en las estimaciones.

En el caso de la HPA, la regla de redondear cada variable al 0.0025 de su media se traduce en el siguiente redondeo efectivo:

Tabla 1: Detalle de Redondeo
HPA a Diciembre 2007

Variable	**Redondeo**
Ingreso imponible	500
Retiros Programados	100
Renta Temporal	1.000
Excedente de libre disposición	5.000
Saldo Fondo A	50.000
Saldo Fondo B	10.000
Saldo Fondo C	10.000
Saldo Fondo D	10.000
Saldo Fondo E	10.000
Bono de Reconocimiento	10.000
Monto prima única traspasada a cía. de seguros	50.000
Comisión porcentual por cotizaciones	10
Comisión porcentual por cambio de AFP	10
Comisión porcentual por Retiro Programado	5
Comisión porcentual por Renta Temporal	10
Comisión porcentual por mantención de saldo	5

[2]En este caso, el redondeo es a los US$100 más cercanos, que en términos porcentuales es de 0.25 % aproximadamente.

Así, por ejemplo, si el ingreso imponible observado es 355.200, el valor redondeado será 355.000, al igual que si el ingreso imponible observado es 354.900. En otras palabras, todos los ingresos imponibles observados en el intervalo [354.750, 355.249] son redondeados al múltiplo de 500 más cercano, es decir a 355.000.

Con el objeto de analizar hasta que punto el redondeo afecta los dos primeros momentos de las variables, se obtienen estimaciones de máxima verosimilitud univariadas para la media y la desviación estándar de cada variable en su versión original y redondeada. Luego se estima estos dos estadísticos para las variables redondeadas corregidas, donde la corrección propuesta por Sheppard (1898) consisten en sustraer $h^2/12$ a la varianza estimada, donde h es el ancho del intervalo de redondeo (la media estimada no necesita correción). Las estimaciones de la media y varianza corregida de las variables redondeadas pueden ser vistas como estimaciones de máxima verosilmilitud aproximadas (MLE) y sus errores estándar pueden ser interpretados como error de muestreo del MLE (ver Heitjan (1989)).

Las tablas 2 y 3 presentan, respectivamente, los MLE para las variables en formato de corte transversal y en formato longitudinal de la HPA 2007. En cada tabla, la primera columna representa los MLE para la variable original no redondeada, la segunda columna las estimaciones para la variable redondeada sin corregir y la tercera para la variable redondeada corregida.

Tanto las medias como las desviaciones estándar estimadas para las variables originales, redondeadas y redondeadas corregidas son exactamente iguales para cada una de las variables contenidas en la tabla 2 . La excepción es el Saldo Fondo E, cuya media de la variable original es estadísticamente distinta a la media de la variable redondeada no corregida. Sin embargo, la corrección de Sheppard revierte esta situación haciendo que la media de la variable corregida sea estadísticamente igual a la de la variable original. Por su parte, la estimación de la desviación estándar no llega a ser subsanada por la correción.

La tabla 3 muestra los MLEs para las variables en formato longitudinal. Se observa que, cuando el redondeo tiene efecto, éste es modesto respecto de las variables originales. Al contrario de lo observado para el Saldo Fondo E, esta vez la corrección de Sheppard no logra acercar los estadísticos de las variables redondeadas a los de las variables originales no redondeadas.

Tabla 2: Estimaciones Maxima Verosimilitud
Variables en Formato de Corte Transversal

		Variable original	Variable redondeada no corregida	Variable redondeada corregida
Saldo Fondo A (en logs)	media	15,79 (0,027)***	15,79 (0,027)***	15,79 (0,027)***
	sd	1,41 (0,008)***	1,41 (0,008)***	1,41 (0,008)***
	N	2.801	2.801	2.801
Saldo Fondo B (en logs)	media	14,06 (0,018)***	14,06 (0,018)***	14,06 (0,018)***
	sd	1,77 (0,005)***	1,77 (0,005)***	1,77 (0,005)***
	N	9.945	9.945	9.945
Saldo Fondo C (en logs)	media	14,25 (0,020)***	14,25 (0,020)***	14,25 (0,020)***
	sd	2,25 (0,005)***	2,25 (0,005)***	2,25 (0,005)***
	N	12.995	12.995	12.995
Saldo Fondo D (en logs)	media	13,65 (0,043)***	13,65 (0,043)***	13,65 (0,043)***
	sd	2,59 (0,009)***	2,59 (0,009)***	2,59 (0,009)***
	N	3.657	3.657	3.657
Saldo Fondo E (en logs)	media	16,41 (0,335)***	15,02 (0,121)***	16,90 (0,333)***
	sd	2,93 (0,087)***	2,02 (0,030)***	2,20 (0,061)***
	N	278	278	278
Bono Reconocimiento (en logs)	media	14,85 (0,031)***	14,85 (0,031)***	14,85 (0,031)***
	sd	1,52 (0,009)***	1,52 (0,009)***	1,52 (0,009)***
	N	2.414	2.414	2.414
Monto prima traspasada a cía. de seguros (en logs)	media	16,12 (0,067)***	16,12 (0,067)***	16,12 (0,067)***
	sd	1,79 (0,018)***	1,79 (0,018)***	1,79 (0,018)***
	N	709	709	709

Errores estándar entre paréntesis. * p<0.1, ** p<0.05, *** p<0.01

Table 3: Etimadores Maxima Verosimilitud
Variables en Formato Longitudinal

		Variable original	Variable redondeada no corregida	Variable redondeada corregida
Ingreso Imponible (en logs)	media	11,41 (0,001)***	11,41 (0,001)***	11,41 (0,001)***
	sd	1,41 (0,000)***	1,41 (0,000)***	1,41 (0,000)***
	N	3.242.240	3.242.240	3.242.240
Retiro Programado (en logs)	media	10,79 (0,004)***	10,80 (0,004)***	10,80 (0,004)***
	sd	1,19 (0,001)***	1,19 (0,001)***	1,19 (0,001)***
	N	92.423	92.423	92.423
Renta Temporal (en logs)	media	12,40 (0,019)***	12,40 (0,019)***	12,40 (0,019)***
	sd	1,19 (0,006)***	1,19 (0,006)***	1,19 (0,006)***
	N	4.061	4.061	4.061
Excedente de Libre Disposición (en logs)	media	13,04 (0,061)***	13,04 (0,061)***	13,04 (0,061)***
	sd	1,99 (0,015)***	1,99 (0,015)***	1,99 (0,015)***
	N	1.080	1.080	1.080
Comisión porcentual por cotizaciones	media	4.378,04 (2.772)***	4.378,57 (2.772)***	4.378,58 (2.772)***
	sd	4.881,82 (0.014)***	4.881,82 (0.014)***	4.890,20 (0.014)***
	N	3.100.566	3.100.566	3.100.566
Comisión porcentual por cambio de AFP	media	8.387,43 (3868.750)**	8.321,13 (3868.511)**	8.628,67 (3870.069)**
	sd	65.609,75 (5.340)***	65.609,24 (5.340)***	65.626,66 (5.341)***
	N	288	288	288
Comisión porcentual por Retiro Programado	media	1.127,85 (10.408)***	1.127,82 (10.408)***	1.127,54 (10.408)***
	sd	1.947,63 (0.083)***	1.947,63 (0.083)***	1.949,66 (0.083)***
	N	35.018	35.018	35.018
Comisión porcentual por Renta Temporal	media	5.004,41 (86.527)***	5.004,90 (86.527)***	5.004,80 (86.526)***
	sd	3.606,24 (0.509)***	3.606,24 (0.509)***	3.614,53 (0.509)***
	N	1.737	1.737	1.737
Comisión porcentual por mantención de saldo	media	422,31 (2.198)***	422,28 (2.198)***	422,24 (2.198)***
	sd	1.033,17 (0.024)***	1.033,24 (0.024)***	1.035,17 (0.024)***
	N	220.977	220.977	220.976

Errores estándar entre paréntesis. * p<0.1, ** p<0.05, *** p<0.01

Estudiamos ahora el efecto de redondear en un cotexto de análisis multivariado. En este caso la corrección consiste en sustraer $h^2/12$ de cada elemento de la diagonal de la matriz de varianzas y covarianzas (ver, por ejemplo, Heitjan (1989) y Dempster and Rubin (1983)). Se estima un

versión simple de la ecuación de Mincer para ilustrar el punto. Debido a que la HPA no contiene la educación de los individuos, se utilizan los datos de la Encuesta de Protección Social, EPS, del año 2006[3] para estimar el log de los ingresos del trabajo (redondeados según la regla antes descrita) en función de sexo, edad y dummies de nivel educacional. Los coeficientes estimados para los ingresos originales no corregidos, redondeados no corregidos y redondeados corregidos se muestran en las columnas 1, 2 y 3, respectivamente, de la tabla 4. Tal como en el caso univariado, el redondeo no tiene efecto alguno en los parámetros estimados. Por otro lado, las diferencias entre los estimadores para la variable redondeada no corregida o la variable redondeada corregida y la variable original son menores que el límite de error de 2 desviaciones estándar.

Tabla 4: Etimaciones de Regresión Multivariada

Variable dependiente: Ln(ingreso imponible)

	Sin redondeo	Con redondeo, sin corrección	Con redondeo, con corrección
constante	11,22	11,22	11,16
	(0,053)***	(0,053)***	
sexo	0,35	0,35	0,36
	(0,0216)***	(0,0216)***	
edad	0,00	0,00	0,00
	(0,001)**	(0,001)**	
educación primaria	0,42	0,42	0,44
	(0,03)***	(0,0306)***	
educación secundaria	0,66	0,66	0,68
	(0,034)***	(0,035)***	
educación universitaria	1,24	1,24	1,26
	(0,035)***	(0,034)***	
N	8.288	8.288	8.288

Errores estándar entre paréntesis. * p<0.1, ** p<0.05, *** p<0.01

5. Conclusiones

La Superintendencia de Pensiones cuenta con valiosas base de datos de los afiliados al sistema de AFP y al Seguro de Cesantía, que contienen información laboral, previsional y de ahorro para el desempleo.

La Superintendencia, en su interés por aumentar el conocimiento y para contar con más herramientas para un debate informado, pone a disposición de la comunidad académica, diseñadores de políticas públicas y todos aquellos que quieran hacer estudios en el área una muestra representativa de los afiliados al Seguro de Cesantía. Sin embargo, existe una tensión natural entre hacer pública esta información y proteger la identidad de los individuos cuyos datos están en las bases.

[3]La EPS es realizada por la Universidad de Chile, por encargo del Ministerio del Trabajo. Más información en www.proteccionsocial.cl

Para resolver esta tensión, la Superintendencia ha adoptado las siguientes medidas que intentan disminuir la probabilidad de identificar a los individuos: se reemplaza el RUT de individuos y empleadores por un identificador único, las fechas contienen mes y año pero omiten día, la información geográfica más desagregada es a nivel comunal y las variables monetarias han sido redondeadas a 0.25 % de su media aproximadamente.

Este artículo provee evidencia del efecto que el redondeo tiene en las medias y desviaciones estádar de las variables. En el caso univariado, el redondeo no tiene ningún efecto para la mayoría de las variables y sólo un efecto insignificante para otras. En la misma linea, el ejercicio multivariado simple que se desarrolla muestra que el redondeo no tiene efectos en las estimaciones. Finalmente, la evaluación de la fórmula de corrección disponible para datos redondeados sugiere que es bastante inocua tanto en el caso univariado como en el multivariado.

Referencias

Dempster, A. and D. Rubin (1983): Rounding Error in Regression: The Appropriatness of Sheppard's Correction. *Journla of the Royal Statistical Society. Series B (Methodological)*, 45(No. 1).

Heitjan, D. (1989): Inference from Grouped Continous Data: A Review. *Statistical Science*, 4(No. 2).

Sheppard, W. (1898): On the calculation if the Most Probable Values of Frecuency Constants for Data Arranged According to Equidistant Divisions of a Scale. *Proc. London Math. Society*, 29.