UNIVERSIDAD SAN FRANCISCO DE QUITO

Colegio de Posgrados

Estimación de Modelos de Regresión en R como una alternativa a Stata

Bolívar Efraín Morales Oñate

Carlos Jiménez Mosquera, Ph.D., Director de Tesis

Trabajo de titulación presentado como requisito

para la obtención del título de Magister en Matemáticas Aplicadas

Quito, 21 de diciembre de 2015

Universidad San Francisco de Quito

Colegio de Posgrados

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Estimación de Modelos de Regresión en R como una alternativa a Stata

Bolívar Efraín Morales Oñate

Carlos Jiménez Mosquera, Ph.D.
Director de Tesis y
Director de la Maestría en Matemáticas Aplicadas _____

Julio César Ibarra Fiallo, M.Sc
Miembro del Comité de Tesis _____

Carlos Jiménez Mosquera, Ph.D.
Miembro del Comité de Tesis _____

César Zambrano, Ph.D.
Decano de la Escuela de Ciencias _____

Hugo Burgos, Ph.D.
Decano del Colegio de Posgrados _____

Quito, 21 de diciembre de 2015

© DERECHOS DE AUTOR

Firma: _____

Nombre: Bolívar Efraín Morales Oñate

C. I.: 1803095858

Fecha: Quito, diciembre de 2015

DEDICATORIA

A Dios por sobre todas las cosas. A mi madre, quien con su esfuerzo me ha apoyado siempre en todo. A mi padre que siempre ha estado ahí con palabras de aliento. A mis hermanos pero en especial a "Tito" (Víctor Morales O.) quien fue mi soporte antes, durante y después de la carrera. A mi hija quien en medio de la tempestad y su tierna edad con un abrazo me ha inyectado toda su energía. A mi esposa quien de todas las maneras posibles también aportó para la realización de este proyecto de vida. Y a todas las personas que de una u otra forma han sumado para la realización de este sueño.

AGRADECIMIENTOS

A Carlos Jiménez, gratitud eterna y mi más sincera admiración.

RESUMEN

Stata y R son dos de los programas más utilizados para abordar problemas de naturaleza econométrica. Este documento pretende mostrar las equivalencias que se pueden tener tanto en un software comercial (Stata) como en uno libre (R) en lo referente al manejo de ciertas regresiones. Representa un primer paso para los usuarios interesados en la estimación de modelos. En cada comando analizado se presente una descripción breve del objetivo del modelo, los datos utilizados y la descripción de resultados.

ABSTRACT

Stata and R are two of the most used programs to address problems of econometric nature. This document aims to show the equivalence that can be both commercial software (Stata) and one free (R) in relation to the handling of certain regressions. It represents a first step for users interested in estimating models. Each analyzed command is presented as a brief description of the purpose of the model, the data and results.

TABLA DE CONTENIDO

TABLAS

FIGURAS

INTRODUCCIÓN

El análisis estadístico, y en particular, la estimación de regresiones lineales y no lineales, ha sido llevado a cabo de manera más eficiente a través del uso de herramientas computacionales. Esto permite realizar la estimación de parámetros a través de procesos que, sin la disponibilidad computacional, serían casi inalcanzables.

El objetivo principal de este trabajo es implementar un conjunto de comandos paralelos a la oferta de Stata para una lista de modelos de regresión (ver tabla 1). Esto se lo realizó de dos maneras: i) se investigó si existía un comando equivalente en R para los diferentes modelos. Y, ii) los comandos que no estaban disponibles en R implementados fueron programados.

En muchos problemas existe una relación inherente entre dos o más variables, y resulta necesario explorar la naturaleza de esta relación. El análisis de regresión es una técnica estadística para el modelado y la investigación de la relación entre dos o más variables[1]. Unas aplicaciones generales serían: líneas de tendencia en series de tiempo, relaciones causales en medicina, modelos econométricos, etc. El análisis de regresión puede emplearse para construir un modelo que permita describir y/o predecir el rendimiento para una temperatura dada.

R es un lenguaje y entorno de programación para análisis estadístico y gráfico. Se trata de un proyecto de software libre, resultado de la implementación GNU del premiado lenguaje S. R y S-Plus -versión comercial de S- son, probablemente, los dos lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy populares en el campo de la investigación biomédica, la bioinformática y las matemáticas financieras. A esto contribuye

[1] Algunos investigadores incluso usan regresión para la identificación de relaciones de causalidad.

la posibilidad de cargar diferentes bibliotecas o paquetes con finalidades específicas de cálculo o gráfico.

Stata es un paquete de software estadístico creado en 1985 por StataCorp. Es utilizado principalmente por instituciones académicas y empresariales dedicadas a la investigación, especialmente en economía, sociología, ciencias políticas, biomedicina y epidemiología.
Stata permite, entre otras funcionalidades, la gestión de datos, el análisis estadístico, el trazado de gráficos y las simulaciones.

Es claro que realizar un desarrollo exhaustivo de cada uno de los modelos y métodos de la tabla 1 es una tarea que rebasa el alcance de este trabajo. Su alcance es un sentido comparativo de carácter inicial. Usuarios que deseen profundizar en los métodos pueden acceder a las referencias mencionadas a lo largo del documento.

A continuación se muestran los comandos de Stata sobre los cuales se desarrolla este trabajo:

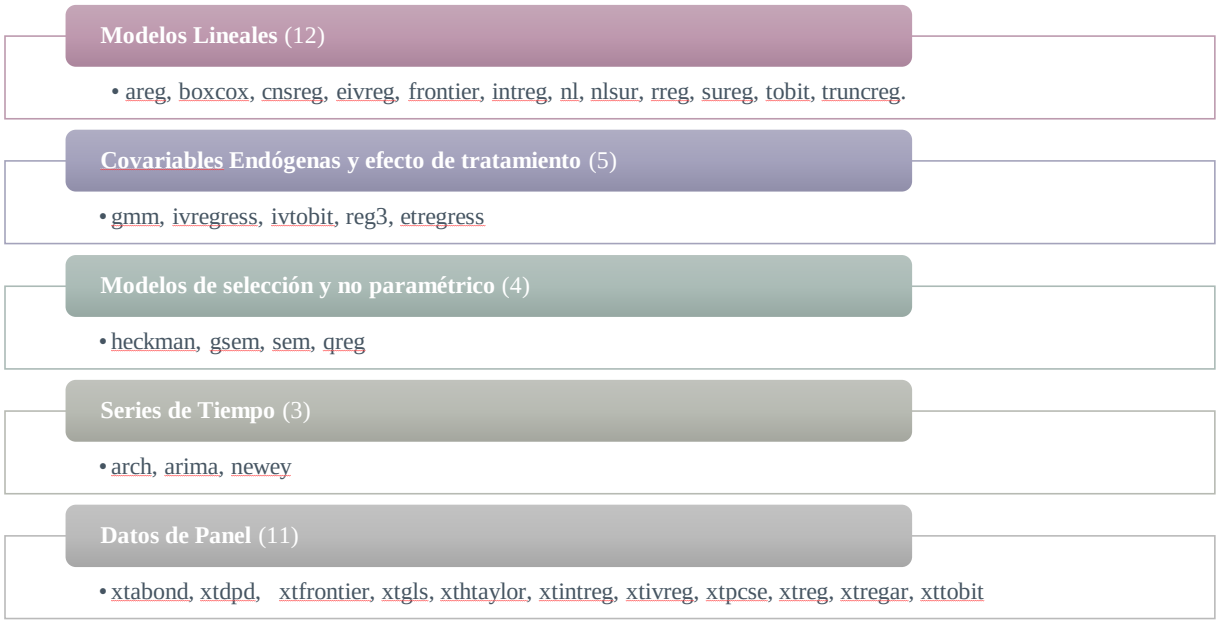
	Equivalente en R		
Stata	**Librería**	**Comando**	**Descripción**
areg	RegUtils	alm	Ajusta regresiones con variables dummy
arch	rugarch	ugarchFit	Modelos de regresión con errores ARCH
arima	fArma	armaFit	Modelos ARIMA
boxcox	RegUtils	boxcox.r	Modelos de regresión Box–Cox
cnsreg	stats	lm	Regresión lineal restringida
eivreg	RegUtils	eivlm	Regresión con errores en las variables
etregress	RegUtils	etreg	Regresión lineal con efectos endógenos
frontier	frontier	sfa	Modelos estocásticos frontier
gmm	gmm	gmm	método generalizado de momentos de estimación
heckman	sampleSelection	selection	Modelos de selección Heckman
intreg	intReg	intReg	Regresión intervalo
ivregress	AER	ivreg	Regresión de variables instrumentales simples
ivtobit	RegUtils	ivtobit	Regresión tobit con variables endógenas
newey	Sandwich	NeweyWest	Regresión de Newey –West con errores estándar
nl	stats	nls	Estimación no lineal de mínimos cuadrados
nlsur	systemfit	nlsystemfit	Estimación de sistemas de ecuaciones no lineales
qreg	quantreg	rq	Regresión cuantil (incluye mediana)
reg3	systemfit	systemfit	Regresión de mínimos cuadrados en tres estados (3SLS)
rreg	MASS	rlm	Un tipo de regresión robusta
gsem	lavaan	cfa	modelos de ecuaciones estructurales generalizadas
sem	lavaan	cfa	modelos de ecuaciones estructurales lineales

sureg	systemfit	systemfit	regresión aparentemente no relacionada
tobit	censReg	censReg	Regresión tobit
truncreg	truncreg	truncreg	regresión truncada
xtabond	plm	pgmm	Estimación de datos de panel dinámicos de Arellano–Bond
xtdpd	plm	pgmm	Estimación de datos de panel dinámicos lineales
xtfrontier	frontier	sfa	Modelos frontier de datos de panel dinámicos
xtgls	panelAR	panelAR	Modelos GLS de datos de panel
xthtaylor	plm	pht	Modelos de estimación de errores de Hausman–Taylor
xtintreg	plm	survreg	Modelos de regresión de datos de panel en intervalos
xtivreg	plm	plm	Regresión de datos de panel con variables instrumentales (2SLS)
xtpcse	panelAR	panelAR	Regresión lineal con errores estándar
xtreg	plm	plm	Modelos lineales con efectos aleatorios
xtregar	nlme	lme	Modelos lineales con efectos aleatorios AR(1)
xttobit	censReg	censReg	Modelos tobit de datos de panel

Tabla 1 Resumen de comandos en Stata

Entre las funciones disponibles en la biblioteca estándar o paquetes de terceros para el software R, no se encontraron comandos que permitan estimar los modelos implementados en los siguientes comandos de Stata: **`areg`**, **`boxcox`**, **`eivreg`**, **`etregress`**, **`ivtobit`** razón por la cual fueron construidos como parte del presente trabajo de titulación y se los presenta en el paquete RegUtils el cual se detalla tanto en los resultados como en los anexos.

La siguiente tabla muestra juna clasficación de los comandos estudiados según lo hace Stata y en función de la cual se desarrollará este trabajo:



La forma en la que se aborda cada instrucción es a través de la descripción de tres componentes:

- Objetivo del modelo y datos usados
- Formulación del modelo
- Estimación y descripción de los resultados.

El uso del software R permite que los usuarios puedan tener un mayor seguimiento del proceso de estimación. Por un lado, el código de programación suele ser más explícito que en los programas comerciales como Stata. Además, en la mayoría de los casos el código fuente de los comandos suele ser oculto para el usuario pero en R se tiene acceso total a estas fuentes. Esto último ayuda a que la persona pueda complementar su aprendizaje del modelo que analiza. Es por esto que otro de los objetivos buscados en este trabajo es mostrar una comparación entre los comandos utilizados para el ajuste de estos modelos en Stata y sus equivalentes en R, señalando las diferencias encontradas entre los dos programas comparados.

Finalmente se presentan conclusiones y recomendaciones acerca de posibles formas de mejorar las herramientas de software libre con respecto a su uso en aplicaciones de estimación de modelos de regresión.

ANÁLISIS Y RESULTADOS

A continuación se muestran los resultados del ajuste de los Modelos de regresión comparando entre la salida de los comandos de R y Stata.

MODELOS LINEALES

Los siguientes 12 modelos han sido agrupados debido a que son variaciones de modelos lineales. Es decir que comparten la característica de tener linealidad en los parámetros.

Regresión Lineal con conjuntos grandes de variables binarias (areg[2])

Descripción

Se desea ajustar un modelo de regresión lineal que tenga factores[3] dentro del conjunto de variables explicativas de modo que el ajuste implique el uso de un gran número de variables binarias.

Para ejemplificar el método, se dispone de un conjunto de datos que describen las características de 74 autos. Las variables de interés son:

Variable	**Descripción**	**Tipo de variable**
Mpg	millas por galón	Numérica
Weight	peso en libras	Numérica
gear_ratio	proporción entre plato y piñón	Numérica
rep78	medida (de 1 al 5) del registro de reparaciones del auto donde 1 es la peor y 5 es la mejor	Factor

Se desea explicar las millas por galón en función de las demás variables expresadas en la tabla anterior. Una forma de resolver el problema es ajustar un modelo de regresión tradicional.

[2] http://www.stata.com/manuals14/rareg.pdf

[3] Variables cualitativas que pueden ser ordinales (que tengan un orden. Por ejemplo: leve, moderado, fuerte) o nominales (que no tengan orden. Por ejemplo: colores, etnia)

Este enfoque generaría 4 niveles para el factor rep78[4]. Sin embargo, areg permite, en lugar de realizar una prueba t para cada coeficiente, hacer una prueba F conjunta para todos los niveles de un determinado factor. Este proceso permite trabajar de forma eficiente sobre conjuntos de datos donde la cantidad de variables binarias generadas es muy grande, pero también nos permite centrar el análisis en un conjunto de variables independientes al margen de los grupos creados por las variables binarias.

Modelo

Suponga que tiene una regresión con un gran número de variables binarias:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{d_1\gamma_1} + \mathbf{d_2\gamma_2} + \cdots + \mathbf{d_k\gamma_k} + \boldsymbol{\epsilon}$$

donde $\mathbf{d_i}$ son las variables binarias. En las estimaciones del ejemplo se tiene:

$$\text{mpg} = 34.058 - 0.0051\text{weight} + 0.901\text{gear}_\text{r}\text{atio},$$

Se aprecia que los niveles de la variable rep78 no se imprimen debido al método utilizado para obtener más eficiencia.

Aplicación

Aquí se ajusta el modelo descrito en la parte anterior tanto en Stata como en R. Las figuras 1 y 2 a continuación se muestran los resultados en R y Stata respectivamente.

[4] En general, si se tiene k niveles en un factor, se ajustan (k-1) coeficientes.

```
. areg mpg weight gear_ratio, absorb(rep78)
Linear regression, absorbing indicators          Number of obs   =         69
                                                 F(   2,     62) =      41.64
                                                 Prob > F        =     0.0000
                                                 R-squared       =     0.6734
                                                 Adj R-squared   =     0.6418
                                                 Root MSE        =     3.5109

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0051031   .0009206    -5.54   0.000    -.0069433   -.003263
  gear_ratio |    .901478   1.565552     0.58   0.567    -2.228015   4.030971
       _cons |   34.05889   7.056383     4.83   0.000     19.95338    48.1644
-------------+----------------------------------------------------------------
       rep78 |               F(4, 62) =     1.117   0.356           (5 categories)
------------------------------------------------------------------------------
```

Figura 1. Regresión Lineal con conjuntos grandes de variables binarias ajustado en Stata

Note que el comando especifica `absorb(rep78)`. Esto indica que se aplicará una prueba F sobre el conjunto de los niveles de la variable rep78. Su p-valor indica que, en conjunto, los niveles de rep78 no son significativos. Es decir que los niveles de rep78 no influyen en las millas por galón. Los coeficientes de las demás variables se ajustan en forma estándar. Por otro lado, en R sería[5]:

```
library("RegUtils")
datos$rep78 = factor(datos$rep78)
summary(alm(mpg~weight+gear_ratio + rep78, data=datos, absorb="rep78"))
## Call:
## alm(formula = mpg ~ weight + gear_ratio + rep78, data = datos,
##     absorb = "rep78")
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.0588892  7.0105697   4.858 8.40e-06 ***
## weight      -0.0051031  0.0009206  -5.544 6.48e-07 ***
## gear_ratio   0.9014780  1.5655516   0.576    0.567
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.511 on 62 degrees of freedom
## Multiple R-squared:  0.6734, Adjusted R-squared:  0.6418
## F-statistic: 63.93 on 2 and 62 DF,  p-value: 8.571e-16
##
##
## Chi-squared Wald Test:
## X2 = 83.3, df = 2, P(> X2) = 0.0
```

Figura 2 Regresión Lineal con conjuntos grandes de variables binarias ajustado en R

Finalmente, los coeficientes para weight, gear_ratio y la constante son exactamente los mismos al igual que el Error estándar, valores t, y consecuentemente el valor p (ver tablas de

[5] Haciendo uso de la función alm del paquete RegUtils.

coeficientes). R no acostumbra reportar intervalos de confianza, pero con el comando **confint**(nombre de la regresión *fit1*) obtienes los mismos valores (con una pequeña diferencia en el intercepto).

El primer estadístico F (igual a 41.64) mostrado en Stata corresponde al Chi-cuadrado del test de Wald (83.28) (mostrado al final en la salida de R), dividido para los grados de libertad (2). En ambos casos el chi-cuadrado es el mismo y lo más importante la conclusión de la prueba es la misma.

También se observa que coinciden el R^2, el R^2 ajustado y los errores estándar residuales (MSE)=3.511, mismos que se suelen utilizar para evaluar la regresión.

El test F para rep78, no se presenta de forma estándar en R, pero se puede calcular mediante el comando **anova** con el que se obtiene el mismo valor 1.117

Modelos de Regresión con transformaciones de Box Cox (boxcox[6])

Descripción

Se desea encontrar por máxima verosimilitud los parámetros de regresión con variables a las cuales se les ha aplicado la transformación de Box-Cox

En el siguiente ejemplo se utilizarán un subconjunto de los datos de la Segunda Encuesta Nacional de Salud y Nutrición (NHANES II), para crear un modelo de estimación del nivel individual de presión arterial. Las variables a utilizar en el modelo son:

Variable	Descripción	Tipo de variable
bpdiast	Presión arterial	Numérica
bmi	Índice de masa corporal	Numérica
tcresult	Colesterol sérico	Numérica
age	Edad	Numérico
sex	sexo	Factor

[6] http://www.stata.com/manuals13/rboxcox.pdf

Para corregir la no linealidad en la relación con las variables, se calculará la transformación de Box-Cox para las variables bmi, y tcresult, así como la transformación de la variable dependiente, dentro del modelo de regresión.

Modelo

Box and Cox (1964) proponen una transformación de los datos que es útil para reducir el sesgo, estabilizar la varianza y hacer que los datos tengan una distribución más parecida a una normal, entre otras. El modelo usado en el ejemplo se conoce como el modelo *theta*:

$$y_j^{(\theta)} = \beta_0 + \beta_1 x_{1j}^{(\lambda)} + \beta_2 x_{2j}^{(\lambda)} + \gamma_1 z_{1j} + \gamma_2 z_{2j} + \epsilon_j$$

donde $\epsilon \sim N(0, \sigma^2)$, x_1, x_2 son transformadas por Box-Cox con parámetro λ. Las variables z_1, z_2 son covariables no transformadas. Usando los resultados del ejemplo se tiene:

$$y_j^{(0.198)} = 5.83 + 0.087\text{bmi}^{(0.638)} + 0.0047\text{tcresult}^{(0.638)} + 0.003\text{age} - 0.105\text{sex}$$

Note que todos los coeficientes son significativos.

Aplicación

Para estimar este modelo se utilizará en R el comando **`boxcox.r`**, parte del paquete RegUtils. En Stata se utilizará el comando **`boxcox`**. Estos comandos permiten estimar varios tipos de modelos que incluyen el modelo lambda, donde se usa el mismo parámetro de transformación sobre todas las variables, modelos theta, que utilizan un parámetro diferente para la variable independiente, y modelos que mezclan variables transformadas y no transformadas.

Los dos comandos permiten especificar un subconjunto de variables independientes a las cuales no se aplicará la transformación. En las fig. 3 y 4 se muestran los resultados.

```r
library(RegUtils)
fit1 = boxcox.r(bpdiast~bmi+tcresult+age+sex, data=datos, noTrans=c("age","sex"))
summary(fit1)
```

```
## Call:
## boxcox.r(formula = bpdiast ~ bmi + tcresult + age + sex, data = datos,
##     noTrans = c("age", "sex"))
##
## Box-Cox Estimates :
## Theta = 0.1988
## Lambda = 0.6383
##
## LR Test:
## X2 = 2515.0, df = 5, P(> X2) = 0.0
##
## Coefficients :
##                              Chi.sq df Chi.sq  Pr(>|t|)
## (intercept)  5.8356e+00          NA         NA        NA
## bmi          8.7203e-02  1.3692e+03          1 < 2.2e-16 ***
## tcresult     4.7339e-03  8.1177e+01          1 < 2.2e-16 ***
## age          3.8111e-03  3.1906e+02          1 < 2.2e-16 ***
## sex         -1.0549e-01  2.4328e+02          1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## ---
## log likehood = -39775.0
```

Figura 3 Regresión con transformaciones de Box Cox ajustado en R

```
                                                Number of obs   =     10351
                                                LR chi2(5)      =   2515.00
Log likelihood = -39774.987                     Prob > chi2     =     0.000
```

bpdiast	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/lambda	.6383286	.1577601	4.05	0.000	.3291245	.9475327
/theta	.1988197	.0454088	4.38	0.000	.1098201	.2878193

Estimates of scale-variant parameters

	Coef.	chi2(df)	P>chi2(df)	df of chi2
Notrans				
age	.003811	319.060	0.000	1
sex	-.1054887	243.284	0.000	1
_cons	5.835555			
Trans				
bmi	.0872041	1369.235	0.000	1
tcresult	.004734	81.177	0.000	1
/sigma	.3348267			

Test H0:	Restricted log likelihood	chi2	Prob > chi2
theta=lambda = -1	-40162.898	775.82	0.000
theta=lambda = 0	-39790.945	31.92	0.000
theta=lambda = 1	-39928.606	307.24	0.000

Figura 4 Regresión con transformaciones de Box Cox ajustado en Stata

Aunque los coeficientes para lambda y theta coinciden, no se pudo determinar cómo Stata realiza la prueba z que se muestra, así que se usó el método por defecto de MaxLik. Aparte de

eso el resto de coeficientes coinciden, así como sus pruebas de hipótesis (que son chi cuadrado). Se presenta además el mismo número de observaciones y la misma estimación para sigma, así como la verosimilitud de -39775. Para el test de Wald se puede usar el comando **`logLik.ratio.test`** para mostrar este en R.

Modelos de regresión lineal con restricciones (cnsreg)[7]

Descripción

Se desea ajustar un modelo de regresión lineal, sujeto a un conjunto de restricciones lineales. Como ejemplo se utilizará la base de datos de la Industria Automotriz de 1978, de la cual se tomarán las siguientes variables:

Variable	Descripción	Tipo de variable
mpg	millas por galón	Numérica
weight	peso en libras	Numérica
price	precio	Numérica

El modelo lineal a ajustar será el siguiente:

$$mpg = \beta_0 + \beta_1 weight + \beta_2 price$$

Donde se desea imponer la siguiente restricción: $\beta_1 = \beta_2$.

Modelo

Para obtener la estimación de los coeficientes de una regresión restringida se puede modificar la lista de las variables independientes. Por ejemplo si se quiere ajustar el modelo:

$$\text{mpg} = \beta_0 + \beta_1 \text{price} + \beta_2 \text{weight} + \text{u}$$

y restringir $\beta_1 = \beta_2$, se puede escribir:

$$\text{mpg} = \beta_0 + \beta_1 (\text{price} + \text{weight}) + \text{u}$$

[7] http://www.stata.com/manuals13/rcnsreg.pdf

y estimar la regresión de mpg sobre `price+weight. El coeficiente de la suma sería el coeficiente restringido de β_1 y β_2. Usando las estimaciones del ejemplo siguiente, se tiene:

$$\text{mpg} = 30.36 - 0.00098\text{price} - 0.00098\text{weight}$$

lo que comprueba que $\beta_1 = \beta_2$.

Aplicación

Stata proporciona un comando especial para este tipo de ajustes, con el fin de simplificar la sintaxis necesaria: **cnsreg.** En R las facilidades que brinda el mismo lenguaje, permiten el ajuste de estos modelos utilizando el comando estándar **lm**.

En las figuras 5 y 6 se muestran los resultados de la estimación del modelo propuesto tanto en R como en Stata.

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/auto.dta")
fit1 = lm(mpg~I(price+weight), data=datos)
summary(fit1)

Call:
lm(formula = mpg ~ I(price + weight), data = datos)

Residuals:
   Min     1Q Median     3Q    Max
-7.046 -3.593 -0.536  1.813 17.977

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       30.367177   1.577958   19.24  < 2e-16 ***
I(price + weight) -0.000987   0.000161   -6.13  4.2e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.72 on 72 degrees of freedom
Multiple R-squared:  0.343,   Adjusted R-squared:  0.334
F-statistic: 37.6 on 1 and 72 DF,  p-value: 4.23e-08
```

Figura 5 Regresión lineal con restricciones ajustado en R

```
. use http://www.stata-press.com/data/r13/auto
(1978 Automobile Data)
. constraint 1 price = weight
. cnsreg mpg price weight, constraint(1)
Constrained linear regression                   Number of obs   =         74
                                                F(  1,     72)  =      37.59
                                                Prob > F        =     0.0000
                                                Root MSE        =     4.7220

 ( 1)  price - weight = 0

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |  -.0009875   .0001611    -6.13   0.000    -.0013086   -.0006664
      weight |  -.0009875   .0001611    -6.13   0.000    -.0013086   -.0006664
       _cons |   30.36718   1.577958    19.24   0.000     27.22158    33.51278
------------------------------------------------------------------------------
```

Figura 6 Regresión lineal con restricciones ajustado en Stata

El modelo estimado en ambos casos se puede resumir como:

$$mpg = 30.36 - 0.001weight - 0.001price$$

El mismo que cumple con la restricción deseada.

Finalmente, se aprecia que los coeficientes tanto de los parámetros como de los test de las variables numéricas de las figuras 5 y 6 coinciden.

Modelos de regresión de errores en las variables (eivreg[8])

Descripción

Este modelo implica el ajuste de regresiones en las cuales una o más variables independientes son medidas con ruido aditivo. El sesgo introducido por este ruido trata de ser compensando mediante el uso de un coeficiente de confiabilidad.

En el siguiente ejemplo se utilizará los datos de la industria automotriz, utilizados en ejemplos anteriores, asumiendo que el peso de los autos fue medido con ruido aditivo que puede ser aproximado por una confiabilidad de 0.85. Bajo este supuesto se realizará un modelo de regresión lineal simple para el precio, utilizando las siguientes variables:

[8] http://www.stata.com/manuals13/reivreg.pdf

Variable	Descripción	Tipo de variable
price	precio	Numérica
weight	peso en libras	Numérica
foreign	Extranjero	Binaria

Modelo

Si una covariable del modelo tiene un error de medición, una regresión *tradicional* no estimaría adecuadamente su efecto. Además, las demás covariables del modelo podrían estar sesgados debido a la presencia en el modelo de esa variable. Se puede ajustar el sesgo si se conoce la *confiabilidad* (reliability):

$$\text{reliability} = 1 - (\text{noise variance})/(\text{total variance})$$

Esto es, dado el modelo $\mathbf{y} = \mathbf{X\beta} + \mathbf{u}$, para alguna variable x_i en $\mathbf{X}$, x_i es observada con error, $x_i = x_i^* + e$, y la noise variance es la varianza de e. La varianza total es la varianza de x_i. En el ejemplo se tienen los siguientes resultados:

$$\text{price} = -8257.01 + 4.319\text{weight} + 4637.3\text{foreign}$$

Los coeficientes habrían sido menores de no haber tomado en cuenta la medición con error de weight. Refiérase a Draper, Smith, and Pownell (1966) para más detalles de este tipo de modelos.

Aplicación

En Stata el comando **`eivreg`** ajusta modelos de regresión con errores en las variables. En R se implementó el comando **`eivlm`** como parte del paquete RegUtils.

En la Figura 7 y Figura 8 se muestran los resultados tanto en R como en Stata para el modelo propuesto.

```
library(RegUtils)
fit2 = eivlm(price~weight+foreign, data=datos, rel = c(0.85,1))
summary(fit2)
```

```
## Call:
## eivlm(formula = price ~ weight + foreign, data = datos, rel = c(0.85,
##     1))
##
## Residuals:
##            [,1]
## Min     -3902.6
## 1Q      -1522.1
## Median   -492.5
## 3Q       1062.5
## Max      5909.6
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8257.0172  1452.0862  -5.686 2.68e-07 ***
## weight          4.3198     0.4314  10.013 3.25e-15 ***
## foreign      4637.3196   624.5362   7.425 1.94e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1774 on 71 degrees of freedom
## Multiple R-squared:  0.6483, Adjusted R-squared:  0.6384
## F-statistic: 65.45 on 2 and 71 DF,  p-value: < 2.2e-16
```

Figura 7 Regresión de errores en las variables ajustado en R

```
. eivreg price weight foreign, r(weight .85)
                    assumed
    variable     reliability

      weight          0.8500
           *          1.0000

Errors-in-variables regression

                                Number of obs =        74
                                F(  2,    71) =     50.37
                                Prob > F      =    0.0000
                                R-squared     =    0.6483
                                Root MSE      = 1773.54

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |     4.31985    .431431    10.01   0.000     3.459601    5.180099
     foreign |     4637.32   624.5362     7.43   0.000      3392.03    5882.609
       _cons |    -8257.017   1452.086    -5.69   0.000    -11152.39   -5361.639
------------------------------------------------------------------------------
```

Figura 8 Regresión de errores en las variables ajustado en Stata

El coeficiente de confiabilidad afecta el coeficiente y el error estándar obtenidos para la variable peso. Además se observa que los coeficientes tanto de los parámetros como de los test de las variables numéricas de las figuras 7 y 8 coinciden.

Modélos Estocáticos de Frontier (frontier[9])

Descripción

Los modelos de frontera de producción estocástica fueron introducidos por Aigner, Lovell, Schmidt (1977) Meesun y Broeck(1977).

En el siguiente ejemplo se reproduce un estudio realizado por Greene (2003, pp505), que utiliza datos publicados originalmente por Zellner y Revankar (1969). En este estudio de la industria de manufactura y transporte, se utilizan observaciones de valor añadido, capital y trabajo para estimar una función de producción Cobb-Douglas.

Las variables a utilizar son las siguiente:

Variable	**Descripción**	**Tipo de variable**
lnv	Transformación logarítmica del valor añadido	Numérica
lnk	Transformación logarítmica del capital	Numérica
lnv	Transformación logarítmica del trabajo	Binaria

Modelo

Este tipo de modelos fueron introducidos por Aigner, Lovell, and Schmidt (1977). Se asume que la producción q_i está determinada por:

$$q_i = f(\mathbf{z_i}, \boldsymbol{\beta})\xi_i \exp(v_i)$$

donde $\mathbf{z_i}$ son las covariables de las que depende la producción, $\boldsymbol{\beta}$ sus respectivos parámetros, ξ_i es el nivel de ineficiencia y $\exp(v_i)$ son choques aleatorios. Note que si no existen choques ni ineficiencia, la producción sería óptima, es decir $q_i = f(\mathbf{z_i}, \boldsymbol{\beta})$.

Tomando logaritmos se tiene:

[9] http://www.stata.com/manuals13/rfrontier.pdf

$$\ln(q_i) = \beta_0 + \sum_{j=1}^{k} \beta_j \ln(z_{ij}) + v_i - u_i$$

donde $u_i = -\ln(\xi_i)$. Se asume que $v_i \sim N(0, \sigma_v)$ y que u_i puede tener distribución exponencial, media-normal o normla truncada. De manera similar se puede tener una formulación para el costo:

$$\ln(c_i) = \beta_0 + \beta_q \ln(q_i) \sum_{j=1}^{k} \beta_j \ln(p_{ij}) + v_i + u_i$$

Note que tanto en el costo como en la producción, el efecto de ineficiencia es quien aumenta el costo o reduce la producción.

En el ejemplo posterior se tiene las siguientes estimaciones:

$$\ln v = 2.081 + 0.2585 \ln k + 0.7802 \ln l$$

Por defecto se asume que u_i tiene una distribución media-normal. Note que el estimador de la varianza del efecto de ineficiencia $\sigma_u = 0.221$ es significativo, el estimador de la varianza del choque aleatorio $\sigma_v = 0.1751$ también lo es. Asimismo se muestran sigma2 ($\sigma_S^2 = \sigma_u + \sigma_v = 0.079$) y el ratio $\lambda = \sigma_u/\sigma_v = 1.264$.

Aplicación

Para estimar el modelo se utilizará el comando **`frontier`** de Stata y el comando **`sfa`** del paquete **`frontier`** en R. En la Figura 9 y Figura 10 se muestran los resultados tanto en R como en Stata.

```
cobbDouglas <- sfa(lnv~lnk+lnl, data = datos)
summary( cobbDouglas )
```

```
## Error Components Frontier (see Battese & Coelli 1992)
## Inefficiency decreases the endogenous variable (as in a production function)
## The dependent variable is logged
## Iterative ML estimation terminated after 9 iterations:
## log likelihood values and parameters of two successive iterations
## are within the tolerance limit
##
## final maximum likelihood estimates
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept) 2.081135   0.282372  7.3702 1.704e-13 ***
## lnk         0.258548   0.099264  2.6046  0.009197 **
## lnl         0.780245   0.120737  6.4623 1.031e-10 ***
## sigmaSq     0.079750   0.042716  1.8670  0.061904 .
## gamma       0.615244   0.385716  1.5951  0.110696
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## log likelihood value: 2.469523
##
## cross-sectional data
## total number of observations = 25
##
## mean efficiency: 0.8466951
```

Figura 9 Modelos Estocáticos de Frontier ajustado en R

```
. frontier lnv lnk lnl
Iteration 0:   log likelihood =  2.3357572
Iteration 1:   log likelihood =  2.4673009
Iteration 2:   log likelihood =  2.4695125
Iteration 3:   log likelihood =  2.4695222
Iteration 4:   log likelihood =  2.4695222
Stoc. frontier normal/half-normal model          Number of obs   =        25
                                                 Wald chi2(2)    =    743.71
Log likelihood =  2.4695222                      Prob > chi2     =    0.0000
```

lnv	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnk	.2585478	.098764	2.62	0.009	.0649738	.4521218
lnl	.7802451	.1199399	6.51	0.000	.5451672	1.015323
_cons	2.081135	.281641	7.39	0.000	1.529128	2.633141
/lnsig2v	-3.48401	.6195353	-5.62	0.000	-4.698277	-2.269743
/lnsig2u	-3.014599	1.11694	-2.70	0.007	-5.203761	-.8254368
sigma_v	.1751688	.0542616			.0954514	.3214633
sigma_u	.2215073	.1237052			.074134	.6618486
sigma2	.0797496	.0426989			-.0039388	.163438
lambda	1.264536	.1678684			.9355204	1.593552

```
Likelihood-ratio test of sigma_u=0: chibar2(01) = 0.43   Prob>=chibar2 = 0.256
. predict double u_h, u
```

Figura 10 Modelos Estocáticos de Frontier ajustado en Stata

Finalmente, se aprecia que los coeficientes tanto de los parámetros como de los test de las variables numéricas de las figuras 9 y 10 coinciden.

Modelo de regresión para intervalos (intreg[10])

Descripción

Es una generalización de los modelos tobit, donde la variable dependiente para cada observación corresponde a un punto, un intervalo, datos censurados por izquierda o censurados por derecha.

En el ejemplo se utilizarán datos del ingreso anual de un grupo de mujeres. Los ingresos se registraron en intervalos, menor a 5000, 5001 – 10000, . . ., 25001 – 30000, 30001 – 40000, 40001 – 50000, y más de 50,000. Adicionalmente se emplearán las siguientes variables:

Variable	Descripción	Tipo de variable
age	Edad	Numérica
Nev_mar	Variable binaria: 1 si nunca estuvo casada	Binaria
Rural	Variable binaria: 1 si habita en una región rural	Binaria
School	Años de escolaridad	Numérica
Tenure	Años de permanencia en el trabajo	Numérica

Modelo

Este modelo tiene los mismos supuestos que el modelo tobit (trabaja con datos censurados). Pero generaliza en concepto al tener la posibilidad de trabajar con los tipos de datos expuestos anteriormente (a un punto, un intervalo, datos censurados por izquierda o censurados por derecha).

En el ejemplo se tendría lo siguiente:

$$\text{wage} = -12.70 + 0.791\text{age} - 0.013\text{age}^2 - 0.207\text{nev}_{\text{mar}}$$

$$-3.043\text{rural} + 1.334\text{school} + 0.800\text{tenure}$$

[10] http://www.stata.com/manuals13/rintreg.pdf

Aplicación

El modelo se ajustó en R y en Stata mediante el comando **`intReg`**. En la Figura 11 y la Figura 12 se muestran los resultados:

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/womenwage.dta")

library(intReg)

wagecat1 = sapply(datos$wagecat, function(x) which(x == unique(datos$wagecat)))
wagecat2 = sapply(datos$wagecat, function(x) which(x == unique(datos$wagecat)))
wagecat1 = c(-Inf, unique(datos$wagecat))[wagecat1]
wagecat2 = c(unique(datos$wagecat)[1:length(unique(datos$wagecat))-1], Inf)[wagecat2]
y = cbind(wagecat1, wagecat2)
fit1 = intReg(y~age +I(age^2)+ nev_mar  + rural + school + tenure, data=datos)
summary(fit1)
## --------------------------------------------
## Interval regression
## Maximum Likelihood estimation
## BHHH maximisation, 30 iterations
## Return code 2: successive function values within tolerance limit
## Log-Likelihood: -856.3329
## 488 observations, 8 free parameters (df = 480)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.703794   6.803773  -1.867 0.062487 .
## age           0.791519   0.487352   1.624 0.105006
## I(age^2)     -0.013264   0.008105  -1.637 0.102385
## nev_mar      -0.207618   0.847006  -0.245 0.806469
## rural        -3.043170   0.842903  -3.610 0.000338 ***
## school        1.334736   0.132795  10.051  < 2e-16 ***
## tenure        0.800081   0.084521   9.466  < 2e-16 ***
## sigma         7.299697         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

Figura 11 Modelo de regresión para intervalos ajustado en R

```
. intreg wage1 wage2 age c.age#c.age nev_mar rural school tenure
Fitting constant-only model:
Iteration 0:   log likelihood = -967.24956
Iteration 1:   log likelihood =  -967.1368
Iteration 2:   log likelihood =  -967.1368
Fitting full model:
Iteration 0:   log likelihood = -856.65324
Iteration 1:   log likelihood = -856.33294
Iteration 2:   log likelihood = -856.33293
Interval regression                             Number of obs   =        488
                                                LR chi2(6)      =     221.61
Log likelihood = -856.33293                     Prob > chi2     =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.7914438	.4433604	1.79	0.074	-.0775265	1.660414
c.age#c.age	-.0132624	.0073028	-1.82	0.069	-.0275757	.0010509
nev_mar	-.2075022	.8119581	-0.26	0.798	-1.798911	1.383906
rural	-3.043044	.7757324	-3.92	0.000	-4.563452	-1.522637
school	1.334721	.1357873	9.83	0.000	1.068583	1.600859
tenure	.8000664	.1045077	7.66	0.000	.5952351	1.004898
_cons	-12.70238	6.367117	-1.99	0.046	-25.1817	-.2230583
/lnsigma	1.987823	.0346543	57.36	0.000	1.919902	2.055744
sigma	7.299626	.2529634			6.82029	7.81265

```
Observation summary:        14  left-censored observations
                             0     uncensored observations
                             6 right-censored observations
                           468       interval observations
```

Figura 12 Modelo de regresión para intervalos ajustado en Stata

El modelo puede ser comparado con otros a través del logaritmo de la verosimilitud

También, se aprecia que los coeficientes tanto de los parámetros como de los test de las variables numéricas de las figuras 11 y 12 coinciden.

Modelos de estimación no lineal de mínimos cuadrados (nl[11])

Descripción

El objetivo de este método es ajustar una función de regresión no lineal mediante el método de mínimos cuadrados.

[11] http://www.stata.com/manuals13/rnl.pdf

En el ejemplo se desea ajustar una función de producción CES (elasticidad de sustitución constante), de la forma:

$$\ln \mathrm{Q_i} = \beta_0 - \frac{1}{\rho}\ln\left\{\delta \mathrm{K_i^{-\rho}} + (1-\delta)\mathrm{L_i^{-\rho}}\right\} + \varepsilon_i$$

El ajuste se realizará sobre datos de producción que incluyen las siguientes variables:

Variable	**Descripción**	**Tipo de variable**
output	Q_i Resultado de la firma	Numérica
Capital	K_i Capital de la firma	Numérica
Labor	L_i Trabajo de la firma	Numérica

Se tomarán como condiciones iniciales $\rho = 1$ y $\delta = 0.5$. Esto asume que trabajo y capital tienen igual impacto sobre los resultados.

Modelo

En el ejemplo se desea ajustar una función de producción CES (sustitución de elasticidad constante), de la forma:

$$\ln(\mathrm{Q_i}) = \beta_0 - \frac{1}{\rho}\ln\{\delta \mathrm{K_i^{-\rho}} + (1-\delta)\mathrm{L_i^{-\rho}}\} + \epsilon_i$$

Note que el parámetro no lineal es ρ. $\ln(\mathrm{Q_i})$ es el logaritmo de la producción de la empresa i; $\mathrm{K_i}$ y $\mathrm{L_i}$ son el uso de capital y fuerza laboral de la firma i respectivamente; y ϵ_i es el término de error.

En el ejemplo se tendría el siguiente resultado:

$$\ln(\mathrm{Q_i}) = 3.792 - \frac{1}{1.386}\ln\{0.482\mathrm{K_i^{-1.386}} + (1-0.482)\mathrm{L_i^{-1.386}}\}$$

La elasticidad de sustitución del modelo CES está dada por $\sigma = \frac{1}{1+\rho} = \frac{1}{1+1.386} = .4189$

Aplicación

Para estimar el modelo se utilizará el comando **nl** de Stata y el comando **nls** de R, del paquete **stats**. En la Figura 13 y Figura 14 se muestran los resultados del ajuste

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/production.dta")
```

```
library(stats)

fit1 <- nls(lnoutput ~ b0 - 1/rho*log(delta*capital^(-1*rho) + (1-delta)*labor^(-1*rho)),
                data = datos,
                start = list(b0=0, rho = 1, delta = 0.5))
summary(fit1)
```

```
##
## Formula: lnoutput ~ b0 - 1/rho * log(delta * capital^(-1 * rho) + (1 -
##     delta) * labor^(-1 * rho))
##
## Parameters:
##       Estimate Std. Error t value Pr(>|t|)
## b0     3.79215    0.09968  38.043  < 2e-16 ***
## rho    1.38697    0.47257   2.935  0.00416 **
## delta  0.48236    0.05198   9.280 4.81e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5502 on 97 degrees of freedom
##
## Number of iterations to convergence: 6
## Achieved convergence tolerance: 5.124e-06
```

Figura 13 Estimación no lineal de mínimos cuadrados ajustado en R

```
. use http://www.stata-press.com/data/r13/production
. nl (lnoutput = {b0} - 1/{rho=1}*ln({delta=0.5}*capital^(-1*{rho}) +
> (1 - {delta})*labor^(-1*{rho})))
(obs = 100)
Iteration 0:  residual SS =  29.38631
Iteration 1:  residual SS =  29.36637
Iteration 2:  residual SS =  29.36583
Iteration 3:  residual SS =  29.36581
Iteration 4:  residual SS =  29.36581
Iteration 5:  residual SS =  29.36581
Iteration 6:  residual SS =  29.36581
Iteration 7:  residual SS =  29.36581
```

Source	SS	df	MS		
Model	91.1449924	2	45.5724962	Number of obs	= 100
Residual	29.3658055	97	.302740263	R-squared	= 0.7563
				Adj R-squared	= 0.7513
Total	120.510798	99	1.21728079	Root MSE	= .5502184
				Res. dev.	= 161.2538

lnoutput	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
/b0	3.792158	.099682	38.04	0.000	3.594316	3.989999
/rho	1.386993	.472584	2.93	0.004	.4490443	2.324941
/delta	.4823616	.0519791	9.28	0.000	.3791975	.5855258

```
Parameter b0 taken as constant term in model & ANOVA table
```

Figura 14 Estimación no lineal de mínimos cuadrados ajustado en Stata

Todos los parámetros del modelo son estadísticamente significativos. Se puede observar que los coeficientes tanto de los parámetros como de los test de las variables numéricas de las figuras 13 y 14 coinciden.

Modelos de estimación de sistemas de ecuaciones no lineales (nlsur[12])

Descripción

Estos modelos se estiman por mínimos cuadrados no lineales generalizados factibles (FGNLS). Aplica cuando la covarianza de los errores no es conocida y el método tradicional de mínimos cuadrados presenta estimaciones consistentes pero ineficientes.

Los datos a utilizar en el ejemplo corresponden a un experimento en el que se colocaron dos tipos estrechamente relacionadas de bacterias en una placa de Petri, registrando cada hora el número de cada tipo de bacteria. Se asumirá que la evolución de las poblaciones puede ajustarse con un modelo de crecimiento exponencial.

Modelo

Se quiere ajusta el sistema de ecuaciones:

$$p_1 = \beta_1 \beta_2^t + u_1$$

$$p_2 = \gamma_1 \gamma_2^t + u_2$$

donde p_1 y p_2 son dos poblaciones de bacterias y t es el tiempo. Este sistema permite que exista correlación entre u_1 y u_2.

Usando los resultados se tendría:

$$p_1 = (0.392)(1.119)^t$$

$$p_2 = (0.509)(1.102)^t$$

Se tiene que β_2 y γ_2 son mayores que 1, lo que indica que las poblaciones de bacterias incrementaron su tamaño en el tiempo.

Aplicación

[12] http://www.stata.com/manuals13/rnlsur.pdf

Para estimar el modelo se utiliza el comando **`nlsur`** de Stata y el comando **`nlsystemfit`** de R, del paquete **`systemfit.`** En la Figura 15 y Figura 16 se muestran los resultados de la estimación.

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/petridish.dta")
```

```
library(systemfit)
source('~/Repos/regresiones/shortSummary.R')
p1.formula <- p1 ~ beta1*beta2^t
p2.formula <- p2 ~ gamma1*gamma2^t
start.values <- c(beta1 = 1, beta2 = 1, gamma1 = 1, gamma2 = 1)
model <- list(p1.formula, p2.formula)
model.sur <- nlsystemfit("SUR", model, start.values, data=datos)
shortsummary(model.sur)
```

```
##
## Method:
## SUR
## Equations:
## p1 ~ beta1 * beta2^t
## p2 ~ gamma1 * gamma2^t
##
## Regression:
##      Obs Params      RMSE df      R.sq    adj.R2
## [1,]  25      2 0.4521619 23 0.9322981 0.9293546
## [2,]  25      2 0.3944521 23 0.9321616 0.9292121
##
## Coefficients:
##         Estimate Std. error t value  Pr(> t)
## beta1  0.393920   0.067059   5.874 4.47e-07 ***
## beta2  1.119422   0.009267 120.799  < 2e-16 ***
## gamma1 0.507812   0.069711   7.285 3.42e-09 ***
## gamma2 1.102444   0.007534 146.335  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## SUR, 2 iterations
```

Figura 15 Estimación de sistemas de ecuaciones no lineales ajustado en R

```
. use http://www.stata-press.com/data/r13/petridish
. nlsur (p1 = {b1}*{b2}^t) (p2 = {g1}*{g2}^t)
(obs = 25)
Calculating NLS estimates...
Iteration 0:  Residual SS =  335.5286
Iteration 1:  Residual SS =  333.8583
Iteration 2:  Residual SS =  219.9233
Iteration 3:  Residual SS =  127.9355
Iteration 4:  Residual SS =  14.86765
Iteration 5:  Residual SS =  8.628459
Iteration 6:  Residual SS =  8.281268
Iteration 7:  Residual SS =   8.28098
Iteration 8:  Residual SS =  8.280979
Iteration 9:  Residual SS =  8.280979
Calculating FGNLS estimates...
Iteration 0:  Scaled RSS =  49.99892
Iteration 1:  Scaled RSS =  49.99892
Iteration 2:  Scaled RSS =  49.99892
FGNLS regression
```

Equation	Obs	Parms	RMSE	R-sq	Constant
1 p1	25	2	.4337019	0.9734*	(none)
2 p2	25	2	.3783479	0.9776*	(none)

* Uncentered R-sq

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/b1	.3926631	.064203	6.12	0.000	.2668275	.5184987
/b2	1.119593	.0088999	125.80	0.000	1.102149	1.137036
/g1	.5090441	.0669495	7.60	0.000	.3778256	.6402626
/g2	1.102315	.0072183	152.71	0.000	1.088167	1.116463

Figura 16 Estimación de sistemas de ecuaciones no lineales ajustado en Stata

De los resultados, se aprecia que los coeficientes tanto de los parámetros como de los test de las variables numéricas de las figuras 15 y 16 coinciden.

Modelos de regresión robusta (rreg[13])

Descripción

El objetivo es ajustar una versión robusta modelos de regresión lineal *tradicional*. El método sirve para que la estimación de los coeficientes sean menos sensibles ante, por ejemplo, la presencia de outliers[14].

Los datos usados para ilustrar el método corresponden a un conjunto de datos que contiene los datos de kilometraje y pesos de 74 coches. Las variables son:

Variable	Descripción	Tipo de variable
mpg	Millas por galón	Numérica
weight	Peso del auto en libras	Numérica
foreign	Tipo de auto (foreign=1, domestic=0)	Factor (binaria)

[13] http://www.stata.com/manuals13/rrreg.pdf
[14] Los outliers se establecen a través de de distnacia de cook, cuando es mayor a 1.

Se desea explicar las millas por galón en función de las demás variables de la tabla anterior. Para ello se podría ajustar un modelo de regresión *tradicional*. En ese caso el coeficiente del peso es -0.0066 y del tipo de auto es -1.65. Veremos adelante cómo cambian estos valores al ajustar la versión robusta.

Modelo

Se omite el modelo debido a que se trata de un método de estimación.

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. En las figuras 17 y 18 a continuación se muestran los resultados:

```
. rreg mpg weight foreign
   Huber iteration 1:  maximum difference in weights = .80280176
   Huber iteration 2:  maximum difference in weights = .2915438
   Huber iteration 3:  maximum difference in weights = .08911171
   Huber iteration 4:  maximum difference in weights = .02697328
Biweight iteration 5:  maximum difference in weights = .29186818
Biweight iteration 6:  maximum difference in weights = .11988101
Biweight iteration 7:  maximum difference in weights = .03315872
Biweight iteration 8:  maximum difference in weights = .00721325
Robust regression                               Number of obs =      74
                                                F(  2,    71) =  168.32
                                                Prob > F      =  0.0000
------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0063976   .0003718   -17.21   0.000    -.007139   -.0056562
     foreign |  -3.182639    .627964    -5.07   0.000    -4.434763   -1.930514
       _cons |   40.64022   1.263841    32.16   0.000     38.1202    43.16025
------------------------------------------------------------------------------
```

Figura 17 Modelos de regresión robusta ajustado en Stata

En la salida del comando de Stata **`rreg`** ofrece una gran variación en la estimación del coeficiente del tipo de auto, ahora es -3.18 y antes era -1.65. De no haber tomado en cuenta la versión robusta, incluso se habría concluido que esta variable no es significativa.

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/auto.dta")
```

```
library(MASS)
fit1 = rlm(mpg~weight+foreign, data=datos, psi = psi.bisquare)
summary(fit1)
```

```
##
## Call: rlm(formula = mpg ~ weight + foreign, data = datos, psi = psi.bisquare)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9492 -1.0794  0.3093  1.4493 16.6821
##
## Coefficients:
##             Value    Std. Error t value
## (Intercept)  40.6094   1.2182    33.3369
## weight       -0.0064   0.0004   -17.8631
## foreign      -3.2317   0.6053    -5.3394
##
## Residual standard error: 1.785 on 71 degrees of freedom
```

Figura 18 Modelos de regresión robusta ajustado en R

En R se realiza el ajuste con el comando **`rlm`** el cual se basa en mínimos cuadrados., se aprecia que los coeficientes tanto de los parámetros como de los test de las variables numéricas de las figuras 17 y 18 son los mismos.

Modelos de regresión aparentemente no relacionada (sureg15)

Descripción

El objetivo es ajustar modelos de regresión que parecen no relacionados debido a que parecen ser estimaciones conjuntas de varios modelos de regresión, cada uno con su propio término de error. Sin embargo las regresiones si se relacionan porque los errores asociados con las variables dependientes podría estar correlacionados.

Los datos usados para ilustrar el método corresponden a un conjunto de datos que contiene los datos de kilometraje y pesos de 74 coches. Las variables son:

Variable	Descripción	Tipo de variable
price	Precio del auto	Numérica
weight	Peso del auto en libras	Numérica
foreign	Tipo de auto (foreign=1, domestic=0)	Factor (binaria)
length	Longitud en pulgadas	

[15] http://www.stata.com/manuals13/rsureg.pdf

Se desea explicar las variables de precio y peso. Para ello se podría ajustar modelos de regresión *tradicional* por separado. Una donde la variable dependiente es el precio y dependa del tipo de auto y la longitud. Luego otra regresión para explicar el peso que dependa de las mismas covariables anteriores. Sin embargo, el comando `sureg` permite hacer el test donde los coeficientes de las covariables son simultáneamente igual a cero.

Modelo

A partir de las variables descritas, por ejemplo se podría suponer el modelo:

$$\text{price} = \beta_0 + \beta_1 \text{foreign} + \beta_2 \text{length} + u_1$$

$$\text{weight} = \gamma_0 + \gamma_1 \text{foreign} + \gamma_2 \text{length} + u_2$$

Se podrían estimar los dos modelos por separado. Sin embargo, al usar este método, se realiza un test para saber si $\beta_1 = \gamma_1 = 0$. Luego del ajuste se tiene:

$$\text{price} = -11621.3 + 2801.14 \text{foreign} + 90.212 \text{length}$$

$$\text{weight} = -2850.25 + -133.67 \text{foreign} + 31.44 \text{length}$$

La diferencia de esta estimación (a haber realizado las regresiones independientemente) radica en que se permite que u_1 y u_2 estén correlacionados.

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. En las figuras 19 y 20 a continuación se muestran los resultados:

```
. use http://www.stata-press.com/data/r13/auto
(1978 Automobile Data)
. sureg (price foreign length) (weight foreign length), small dfk
Seemingly unrelated regression
```

Equation	Obs	Parms	RMSE	"R-sq"	F-Stat	P
price	74	2	2474.593	0.3154	16.35	0.0000
weight	74	2	250.2515	0.8992	316.54	0.0000

	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
price						
foreign	2801.143	766.117	3.66	0.000	1286.674	4315.611
length	90.21239	15.83368	5.70	0.000	58.91219	121.5126
_cons	-11621.35	3124.436	-3.72	0.000	-17797.77	-5444.93
weight						
foreign	-133.6775	77.47615	-1.73	0.087	-286.8332	19.4782
length	31.44455	1.601234	19.64	0.000	28.27921	34.60989
_cons	-2850.25	315.9691	-9.02	0.000	-3474.861	-2225.639

Figura 19: Regresión aparentemente no relacionada ajustado en Stata

El comando de Stata **`sureg`** permite que, a diferencia de ajustar dos regresiones separadas, que los errores de la regresión estén correlacionados para que las estimaciones sus errores sean más *reales* o tomen en cuenta esta correlación.

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/auto.dta")

library(systemfit)

eq1 <- price ~ foreign + length
eq2 <- weight ~ foreign + length
eqSystem <- list(eqPrice = eq1, eqWeight = eq2)
fitsur <- systemfit(eqSystem, method="SUR", data=datos)
summary(fitsur)
```

```
## SUR estimates for 'eqPrice' (equation 1)
## Model Formula: price ~ foreign + length
##
##               Estimate  Std. Error  t value   Pr(>|t|)
## (Intercept) -11621.3495   3124.4362 -3.71950 0.00039570 ***
## foreign       2801.1429    766.1170  3.65629 0.00048728 ***
## length          90.2124     15.8337  5.69750 2.5665e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2474.593341 on 71 degrees of freedom
## Number of observations: 74 Degrees of Freedom: 71
## SSR: 434776466.563372 MSE: 6123612.205118 Root MSE: 2474.593341
## Multiple R-Squared: 0.315383 Adjusted R-Squared: 0.296098
##
##
## SUR estimates for 'eqWeight' (equation 2)
## Model Formula: weight ~ foreign + length
##
##               Estimate  Std. Error  t value   Pr(>|t|)
## (Intercept) -2850.24968   315.96906 -9.02066 2.1494e-13 ***
## foreign      -133.67750    77.47615 -1.72540   0.088805 .
## length         31.44455     1.60123 19.63770 < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 250.251527 on 71 degrees of freedom
## Number of observations: 74 Degrees of Freedom: 71
## SSR: 4446433.704653 MSE: 62625.826826 Root MSE: 250.251527
## Multiple R-Squared: 0.899161 Adjusted R-Squared: 0.89632
```

Figura 20: Regresión aparentemente no relacionada ajustado en R

En R se utilizó **`systemfit`** con unas adecuaciones en el argumento el mismo que se basa en el método de mínimos cuadrados. De lo que se ve en los resultados tanto en R como en Stata se despliegan los mismos resultados en sus coeficientes y tests.

Modelos de regresión tobit (tobit[16])

Descripción

El objetivo es ajustar modelos de regresión donde las variables dependientes están censuradas[17]. Es decir que solo se dispone de parte de la información de la variable dependiente.

Los datos usados para ilustrar el método corresponden a un conjunto de datos que contiene los datos de kilometraje y pesos de 74 coches. Las variables son:

Variable	Descripción	Tipo de variable
Mpg	millas por galón	Numérica
Wgt	Peso en libras / 1000	Numérica

Se desea explicar las millas por galón en función de las demás variables expresadas en la tabla anterior. Esto se resolvería usando regresión lineal tradicional. Pero se considera que la variable de millas por galón está censurada. `mpg` tiene valores entre 12 y 41 y se considera que no se ha observado el número de millas menor o igual a 17.

Modelo

Con la particularidad de que se censuraron los datos (mpg) con el criterio de que el valor mínimo sea 17 (14 valores censurados), es decir que son censurados por la izquierda. El modelo ajustado sería:

$$\text{mpg} = 41.49 - 6.87\text{wgt}$$

Note que se reporta un valor de sigma. Este valor es comparable con la raíz del error medio cuadrático de un modelo de regresión sin censura en los datos.

[16] http://www.stata.com/manuals13/rtobit.pdf

[17] Una muestra censurada se presenta cuando en una muestra la información sobre la variable dependiente está disponible sólo para algunas observaciones. Debe diferenciarse de una muestra truncada, en la cual la información sobre las variables independientes sólo está disponible si se observa la variable dependiente

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. En las figuras 21 y 22 a continuación se muestran los resultados:

```
. replace mpg=17 if mpg<=17
(14 real changes made)
. tobit mpg wgt, ll
Tobit regression                                  Number of obs     =         74
                                                  LR chi2(1)        =      72.85
                                                  Prob > chi2       =     0.0000
Log likelihood = -164.25438                       Pseudo R2         =     0.1815

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wgt |  -6.87305   .7002559    -9.82   0.000    -8.268658   -5.477442
       _cons |  41.49856   2.05838     20.16   0.000     37.39621    45.6009
-------------+----------------------------------------------------------------
       /sigma |  3.845701   .3663309                      3.115605    4.575797
------------------------------------------------------------------------------
  Obs. summary:         18  left-censored observations at mpg<=17
                        56     uncensored observations
                         0 right-censored observations
```

Figura 21: Modelos de regresión tobit ajustado en Stata

El comando de Stata **`tobit`** ajusta este modelo. Note que el comando replace `mpg=17 if mpg<=17` crea la variable con censura (18 autos). Además se tiene el número de observaciones sin censura, 56.

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/auto.dta")
datos$cmpg = datos$mpg
datos$cmpg[datos$cmpg<17] = 17
datos$wgt = datos$weight/1000

library(censReg)

fit1 = censReg(cmpg~wgt, data=datos, left=17)
summary(fit1)
## Coefficients:
##             Estimate Std. error t value Pr(> t)
## (Intercept) 41.49856    2.05838  20.161  <2e-16 ***
## wgt         -6.87305    0.70026  -9.815  <2e-16 ***
## logSigma     1.34696    0.09526  14.140  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Newton-Raphson maximisation, 5 iterations
## Return code 1: gradient close to zero
## Log-likelihood: -164.2544 on 3 Df
```

Figura 22: Modelos de regresión tobit ajustado en R

En R se utilizó el comando **`censReg`** para ajustar el modelo[18]. Se puede observar que los resultados en R y Stata coinciden en sus coeficientes y sus parámetros.

Modelos de regresión truncada (truncreg[19])

Descripción

El objetivo es ajustar modelos de regresión de variables dependientes en independientes que provienen de la realización de una muestra de una parte restringida de toda población. Se asume que, bajo el supuesto de normalidad para toda la población, los errores siguen una distribución normal truncada. Es decir que se establece un umbral para la variable dependiente.

Los datos usados para ilustrar el método corresponden a un conjunto de datos que contiene 753 observaciones sobre la oferta de trabajo de las mujeres. La submuestra es de 250 observaciones, con 150 trabajadores de mercado y no de mercado 100 trabajadores. Las variables son:

Variable	**Descripción**	**Tipo de variable**
lfp	Igual a 1 si a esposa trabajó en 1975	Factor (Binaria)
whrs	Horas de trabajo de la esposa	Numérica
kl6	Número de hijos menores a 6 años	Numérica
k618	Número de hijos entre 6 y 18 años	Numérica
wa	Edad de la esposa	Numérica
we	Escolaridad de la esposa	Numérica

Se desea estimar las horas de trabajo en función de las demás variables de la tabla anterior. La opción que permite truncar es `ll(0)`.

[18] el cual se basa en máxima verosimilitud
[19] http://www.stata.com/manuals13/rtruncreg.pdf

Modelo

Usando los datos descritos anteriormente, el modelo estimado sería:

$$\text{whrs} = 1586.26 - 803.00\text{kl6} - 172.875\text{k618} - 8.82\text{wa} + 16.52\text{we}$$

Note que los datos de este modelo fueron censurados si son menores a 0.

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. En las figuras 23 y 24 a continuación se muestran los resultados:

```
. truncreg whrs kl6 k618 wa we, ll(0)
(note: 100 obs. truncated)
Fitting full model:
Iteration 0:   log likelihood = -1205.6992
Iteration 1:   log likelihood = -1200.9873
Iteration 2:   log likelihood = -1200.9159
Iteration 3:   log likelihood = -1200.9157
Iteration 4:   log likelihood = -1200.9157
Truncated regression
Limit:    lower =          0                      Number of obs =    150
          upper =       +inf                      Wald chi2(4)  =  10.05
Log likelihood = -1200.9157                       Prob > chi2   = 0.0395
------------------------------------------------------------------------------
        whrs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         kl6 |  -803.0042   321.3614    -2.50   0.012    -1432.861   -173.1474
        k618 |   -172.875   88.72898    -1.95   0.051    -346.7806    1.030579
          wa |  -8.821123   14.36848    -0.61   0.539    -36.98283    19.34059
          we |   16.52873   46.50375     0.36   0.722    -74.61695    107.6744
       _cons |   1586.26     912.355     1.74   0.082    -201.9233    3374.442
-------------+----------------------------------------------------------------
       /sigma |   983.7262   94.44303    10.42   0.000     798.6213    1168.831
------------------------------------------------------------------------------
```

Figura 23: Regresión truncada ajustado en Stata

El comando de Stata **`truncreg`** ajusta un modelo truncado. En particular, se trunca la variable dependiente (horas de trabajo de la esposa) de modo que sea mayor a 0. Esto implica que solo se toma en cuenta a las esposas que trabajaron en 1975 asumiendo que los errores siguen una distribución normal truncada.

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/laborsub.dta")
```

We use truncreg to perform truncated regression with truncation from below zero:

```
library(truncreg)
fit1 <- truncreg(whrs~ kl6 + k618 + wa + we, data = datos,
                 subset = whrs>0, point = 0, direction = "left",
                 method="BHHH")
summary(fit1)
```

```
##
## Call:
## truncreg(formula = whrs ~ kl6 + k618 + wa + we, data = datos,
##     subset = whrs > 0, point = 0, direction = "left", method = "BHHH")
##
##
## Coefficients :
##              Estimate Std. Error t-value  Pr(>|t|)
## (Intercept) 1586.2446   978.0947  1.6218  0.104853
## kl6         -802.9963   289.5257 -2.7735  0.005546 **
## k618        -172.8733   106.9201 -1.6168  0.105912
## wa            -8.8208    15.6446 -0.5638  0.572875
## we            16.5288    53.8797  0.3068  0.759017
## sigma        983.7199    66.8566 14.7139 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -1200.9 on 6 Df
```

Figura 24: Regresión truncada ajustado en R

En R el nombre que se utilizó tiene el mismo nombre y se basa en el método de máxima verosimilitud. La pequeña diferencia en los resultados que se ven en los coeficientes es por el método de optimización que utiliza cada método de R y Stata.

COVARIABLES ENDÓGENAS Y EFECTO DE TRATAMIENTO

El siguiente grupo de modelos se usan para el manejo de variables endógenas de modo que se mejore su especificación. El último de ellos es un modelo de efecto de tratamiento.

Estimación de modelos por el método generalizado de momentos (gmm[20])

Descripción

El punto de partida del estimador GMM es el principio de analogía. el cual dice que podemos estimar un parámetro reemplazando una condición del momento poblacional con su análogo muestral.

[20] http://www.stata.com/manuals13/rgmm.pdf

La regresión lineal por mínimos cuadrados ordinarios es un estimador de momentos. En el siguiente ejemplo se estimará una regresión lineal utilizando el método de momentos sobre los datos de la industria automotriz. Las variables a utilizar serán las siguientes

Variable	**Descripción**	**Tipo de variable**
mpg	millas por galón	Numérica
weight	peso en libras	Numérica
lenght	longitud	Numérica

Modelo

En un modelo de regresión:

$$y_i = x_i'\beta_0 + u_i$$

donde x_i es el vector de p covariables, β_0 es el verdadero valor de p parámetros desconocidos en β , y u_i es un término de error. En este caso, los momentos poblacionales son $E[g(x_i, \theta_0)] = 0$ que se traducen a $E[x_i u_i] = E[x_i(y_i - x_i'\beta_0)] = 0$. Note que en este caso $g(x_i, \theta) = x_i u_i$ dado que para que el estimador de mínimos cuadrados sea insesgado y consistente se necesita que $E[x_i, u_i] = 0$. Los momentos muestrales están dados por:

$$\frac{1}{n}\sum_{i=1}^{n} x_i \hat{u}_i = \frac{1}{n}\sum_{i=1}^{n} x_i (y_i - x_i'\hat{\beta}) = 0.$$

Entonces el estimador MM (que es un caso particular del GMM) β_0 está dado por $\hat{\beta}$ que resuelve el sistema de p ecuaciones lineales y es equivalente al estimador de mínimos cuadrados ordinarios.

En el ejemplo siguiente se tiene:

$$\text{mpg} = 47.88 - 0.0038\text{weight} - 0.079\text{length}$$

Aplicación

Para estimar el modelo se utilizará el gmm de Stata y R. En la figura 25 y 26 se muestran los resultados obtenidos para los datos del ejemplo.

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/auto.dta")

library(gmm)

## Loading required package: sandwich

y = datos$mpg
x = as.matrix(datos[,c("weight","length")])
summary(gmm(y-x,x))

## Method:  twoStep
##
## Kernel:  Quadratic Spectral
##
## Coefficients:
##              Estimate      Std. Error    t value       Pr(>|t|)
## (Intercept)   4.7885e+01    7.8599e+00    6.0923e+00    1.1127e-09
## xweight      -3.8515e-03    1.9269e-03   -1.9988e+00    4.5627e-02
## xlength      -7.9593e-02    7.0331e-02   -1.1317e+00    2.5776e-01
##
## J-Test: degrees of freedom is 0
##                    J-test                  P-value
## Test E(g)=0:    7.19142477023652e-22   *******
```

Figura 25: Estimación de modelos por el método generalizado de momentos ajustado en R

```
. use http://www.stata-press.com/data/r13/auto
(1978 Automobile Data)
. gmm (mpg - {b1}*weight - {b2}*length - {b0}), instruments(weight length)
Step 1
Iteration 0:   GMM criterion Q(b) =   475.4138
Iteration 1:   GMM criterion Q(b) =  3.305e-20
Iteration 2:   GMM criterion Q(b) =  3.795e-27

Step 2
Iteration 0:   GMM criterion Q(b) =  7.401e-28
Iteration 1:   GMM criterion Q(b) =  3.771e-31

GMM estimation

Number of parameters =   3
Number of moments    =   3
Initial weight matrix: Unadjusted                 Number of obs   =        74
GMM weight matrix:     Robust

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          /b1 |   -.0038515   .0019472    -1.98   0.048    -.0076678   -.0000351
          /b2 |   -.0795935   .0677528    -1.17   0.240    -.2123864    .0531995
          /b0 |    47.88487   7.505985     6.38   0.000     33.17341    62.59633
------------------------------------------------------------------------------
Instruments for equation 1: weight length _cons
```

Figura 26: Estimación de modelos por el método generalizado de momentos ajustado en Stata

Los errores estándar reportados por el comando gmm suelen ser menores que los reportados por los comandos de regresión por mínimos cuadrados, por lo demás los valores son los mismos.

Regresión para variables instrumentales en una sola ecuación (ivregress[21])

Descripción

Estos modelos incluyen regresiones con variables instrumentales y variables instrumentales ponderadas.

En el siguiente ejemplo se emplearán datos del censo de 1980 de vivienda de Estados Unidos, para estimar un modelo para la mediana de la renta mensual por Estado. Las variables a utilizar para el modelo son las siguientes:

Variable	Descripción	Tipo de variable
rent	Mediana de la renta mensual	Numérica
hsngval	Mediana del valor de las viviendas	Numérica
pcturban	Porcentaje de vivienda urbana en el Estado	Numérica

Debido a que los choques aleatorios que afectan las tasas de renta en un estado, probablemente también afectan el valor de una vivienda, en el modelo se tratará la variable hsngval como endógena.

Modelo

El modelo es:

$$y_i = \mathbf{y_i}\boldsymbol{\beta_1} + \mathbf{x_{1i}}\boldsymbol{\beta_2} + u_i$$

$$\mathbf{y_i} = \mathbf{x_{1i}}\,\Gamma_1 \quad + \mathbf{x_{2i}}\,\Gamma_2 \quad + \mathbf{v_i}$$

donde y_i es la variable dependiente para la observación i, $\mathbf{y_i}$ representa a las regresoras endógenas, $\mathbf{x_{1i}}$ representa las regresoras exógenas incluidas y $\mathbf{x_{2i}}$ representa las regresoras

[21] http://www.stata.com/manuals13/rivregress.pdf

exógenas excluidas. $\mathbf{x_{1i}}$ y $\mathbf{x_{2i}}$ en conjunto se llaman instrumentos. u_i y $\mathbf{v_i}$ son términos de error con media cero, y su correlación es cero.

En el ejemplo se estima el modelo:

$$\text{renti} = \beta_0 + \beta_1 \text{hsngvali} + \beta_2 \text{pcturbani} + \text{ui}$$

Se tiene entonces que:

$$\text{renti} = 120.70 + 0.0022 \text{hsngvali} + 0.0815 \text{pcturbani}$$

Los estados con precios de casa más altos tienen rentas más altas. El porcentaje urbano no es significativo.

Aplicación

Para estimar el modelo se utilizará el comando **ivregress** de Stata y el comando **ivreg** de R del paquete AER

En la figura 27 y 28 se muestran los resultados del ejemplo.

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/hsng.dta")

library(AER)

fit1 <- ivreg(rent ~ pcturban + hsngval | pcturban + faminc + as.factor(region),
  data = datos)
summary(fit1)

##
## Call:
## ivreg(formula = rent ~ pcturban + hsngval | pcturban + faminc +
##     as.factor(region), data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -84.1948 -11.6023  -0.5239   8.6583  73.6130
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.207e+02  1.571e+01   7.685 7.55e-10 ***
## pcturban    8.152e-02  3.082e-01   0.265    0.793
## hsngval     2.240e-03  3.388e-04   6.612 3.17e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.86 on 47 degrees of freedom
## Multiple R-Squared: 0.5989,  Adjusted R-squared: 0.5818
## Wald test: 42.66 on 2 and 47 DF,  p-value: 2.731e-11
```

Figura 27: Regresión para variables instrumentales en una sola ecuación ajustado en R

```
. use http://www.stata-press.com/data/r13/hsng
(1980 Census housing data)
. ivregress 2sls rent pcturban (hsngval = faminc i.region)
Instrumental variables (2SLS) regression          Number of obs =      50
                                                  Wald chi2(2)  =   90.76
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =  0.5989
                                                  Root MSE      =  22.166
------------------------------------------------------------------------------
        rent |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hsngval |   .0022398   .0003284     6.82   0.000     .0015961    .0028836
    pcturban |    .081516   .2987652     0.27   0.785    -.504053     .667085
       _cons |   120.7065   15.22839     7.93   0.000     90.85942    150.5536
------------------------------------------------------------------------------
Instrumented:  hsngval
Instruments:   pcturban faminc 2.region 3.region 4.region
```

Figura 28: Regresión para variables instrumentales en una sola ecuación ajustado en Stata

Como se esperaría, los estados con valores más altos de las viviendas tienen las rentas más altas. La proporción de población urbana no parece tener un efecto significativo sobre las rentas.

De lo que se ve los coeficientes tanto de los parámetros como de los test de las variables numéricas de las figuras 27 y 28 coinciden.

Modelo de regresión tobit con variables endogénas (ivtobit[22])

Descripción

Corresponde a modelos tobit donde una o más de las variables regresoras son determinadas endógenamente.

En el ejemplo se utilizarán datos de ingresos de un grupo de mujeres asumiendo que todas las mujeres que deciden no trabajar reciben $ 10,000 en pagos de asistencia social y la manutención de los hijos. En el modelo se incluirán las siguientes variables:

Variable	Descripción	Tipo de variable
Fem_inc	Ingreso	Numérica
Fem_educ	Años de instrucción formal	Numérica
kids	Número de hijos	Numérica
Other_inc	Otros ingresos del hogar	Numérica

[22] http://www.stata.com/manuals13/rivtobit.pdf

Male_educ	Años de instrucción formal del esposo	Numérica

En el modelo se considera que la varibale `Other_inc` es endógena, razón por la cual se utilizará la variable `Male_educ` como variable instrumental.

Modelo

El modelo es:

$$y_{1i}^{*} = \mathbf{y_{2i}}\boldsymbol{\beta} + \mathbf{x_{1i}}\boldsymbol{\lambda} + u_i$$

$$\mathbf{y_{2i}} = \mathbf{x_{1i}}\,\Gamma_1 \quad + \mathbf{x_{2i}}\,\Gamma_2 \quad + \mathbf{v_i}$$

donde $i = 1, \ldots, N$; $\mathbf{y_{2i}}$ es un vector $1 \times p$ de variables endógenas; $\mathbf{x_{1i}}$ es un vector $1 \times k_1$ de variables exógenas; $\mathbf{x_{2i}}$ es un vector $1 \times k_2$ de instrumentos adicionales; y la ecuación $\mathbf{y_{2i}}$ está escrita en forma reducida. El modelo supone que $(u_i, \mathbf{v_i}) \sim N(\mathbf{0})$. $\boldsymbol{\beta}$ y $\boldsymbol{\lambda}$ son vectores de parámetros instrumentales, y Γ_1 y Γ_2 son matrices de parámetros de forma reducida. y_{1i}^{*} es observada en forma censurada.

En el ejemplo se estima el modelo:

$$\text{fem\_inc} = 19.24 - .9045\text{other}_i\text{nc} + 3.27\text{fem\_educ} - 3.31\text{kids}$$

Se aprecia que **v** es significativo por lo que el modelo propuesto podría ser utilizado. Note que la variable exógena es other_inc.

Aplicación

Para estimar el modelo se utilizará el comando **ivtobit** de Stata. En R se utilizará el comando del mismo nombre implementado como parte del paquete RegUtils.

En la figura 29 y 30 se muestran los resultados del ejemplo.

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/laborsup.dta")
```

```
library(RegUtils)
fit1 = ivtobit(fem_inc~other_inc+fem_educ+kids|male_educ+fem_educ+kids,
               data=datos, left = 10)
summary(fit1)
```

```
##
## Call:
## ivtobit(formula = fem_inc ~ other_inc + fem_educ + kids | male_educ +
##     fem_educ + kids, data = datos, left = 10)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##            500            272            228              0
##
## Coefficients:
##             Estimate Std. error t value  Pr(> t)
## (Intercept)  19.2474     6.9470   2.771   0.0056 **
## other_inc    -0.9045     0.1309  -6.912 4.78e-12 ***
## fem_educ      3.2724     0.4169   7.849 4.20e-15 ***
## kids         -3.3124     0.6777  -4.887 1.02e-06 ***
## sigma_u      18.3594     1.1328  16.207  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## BHHH maximisation, 10 iterations
## Return code 2: successive function values within tolerance limit
## Log-likelihood: -3226.085 on 7 Df
## alpha:  0.2907667
## Sigma:
##           [,1]     [,2]
## [1,] 337.0659  80.7641
## [2,]  80.7641 277.7626
```

Figura 29: Regresión tobit con variables endogénas ajustado en R

```
. use http://www.stata-press.com/data/r13/laborsup
. ivtobit fem_inc| fem_educ kids (other_inc = male_educ), ll
Fitting exogenous tobit model
Fitting full model
Iteration 0:   log likelihood = -3228.4224
Iteration 1:   log likelihood = -3226.2882
Iteration 2:   log likelihood =  -3226.085
Iteration 3:   log likelihood = -3226.0845
Iteration 4:   log likelihood = -3226.0845
Tobit model with endogenous regressors          Number of obs   =        500
                                                Wald chi2(3)    =     117.42
Log likelihood = -3226.0845                     Prob > chi2     =     0.0000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   other_inc |  -.9045399   .1329762    -6.80   0.000    -1.165168   -.6439114
    fem_educ |   3.272391   .3968708     8.25   0.000     2.494538    4.050243
        kids |  -3.312357   .7218628    -4.59   0.000    -4.727182   -1.897532
       _cons |   19.24735   7.372391     2.61   0.009     4.797725    33.69697
-------------+----------------------------------------------------------------
      /alpha |   .2907654   .1379965     2.11   0.035     .0202972    .5612336
        /lns |   2.874031   .0506672    56.72   0.000     2.774725    2.973337
        /lnv |   2.813383   .0316228    88.97   0.000     2.751404    2.875363
-------------+----------------------------------------------------------------
           s |   17.70826    .897228                      16.03422    19.55707
           v |   16.66621   .5270318                      15.66461    17.73186
------------------------------------------------------------------------------
Instrumented:  other_inc
Instruments:   fem_educ kids male_educ
------------------------------------------------------------------------------
Wald test of exogeneity (/alpha = 0): chi2(1) =     4.44  Prob > chi2 = 0.0351
  Obs. summary:        272  left-censored observations at fem_inc<=10
                       228     uncensored observations
                         0 right-censored observations
```

Figura 30: Regresión tobit con variables endogénas ajustado en Stata

El estimador utilizado por defecto en ambos comandos es por máxima verosimilitud. Todas las variables del modelo son estadísticamente significativas y se aprecia que son los mismos valores.

Regresiones de mínimos cuadrados de tres estados (reg3[23])

Descripción

El objetivo es estimar sistemas de ecuaciones estructurales con variables endógenas. Generalmente, estas variables endógenas son dependientes de otras ecuaciones en el sistema,

[23] http://www.stata.com/manuals13/rreg3.pdf

pero no siempre. Además este comando se basa en el análisis de variables instrumentales para producir estimaciones consistentes ante la presencia de correlación en el error[24].

Los datos usados para ilustrar el método corresponden a un conjunto de variables maroeconómicas de Estados Unidos. Las variables son:

Variable	**Descripción**	**Tipo de variable**
consump	Consumo	Numérica
wagepriv	Salario privado	Numérica
wagegovt	Salario del gobierno	Numérica
govt	Gasto del gobierno	Numérica
capital1	Valor rezagado del stock de capital	Numérica

El ejemplo se ilustra modelo macroeconómico sencillo que relaciona el consumo de los salarios privados y gubernamentales. Así

$$consump = \beta_0 + \beta_1 wagepriv + \beta_2 wagegovt + \epsilon_1$$

$$wagepriv = \beta_3 + \beta_4 consump + \beta_5 govt + \beta_6 capital1 + \epsilon_2$$

Entonces, `consump` y `wagepriv` serían variables endógenas y `wagegovt`, `govt,` y `capital1` serían exógenas.

Modelo

$$\text{consump} = 19.35 + 0.801\text{wagepriv} + 1.029\text{wagegovt}$$

$$\text{wagepriv} = 14.63 + 0.402\text{consump} + 1.177\text{govt} - 0.028\text{capital1}$$

Se trata de un método de estimación de los parámetros, no de un modelo como tal.

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. En las figuras 31 y 32 a continuación se muestran los resultados:

[24] Se espera que el error esté correlacionado entre las ecuaciones del sistema. Es por esto que el uso de variables instrumentales juega un papel imporante.

```
. use http://www.stata-press.com/data/r13/klein
. reg3 (consump wagepriv wagegovt) (wagepriv consump govt capital1)
Three-stage least-squares regression
```

Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
consump	22	2	1.776297	0.9388	208.02	0.0000
wagepriv	22	3	2.372443	0.8542	80.04	0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
consump						
wagepriv	.8012754	.1279329	6.26	0.000	.5505314	1.052019
wagegovt	1.029531	.3048424	3.38	0.001	.432051	1.627011
_cons	19.3559	3.583772	5.40	0.000	12.33184	26.37996
wagepriv						
consump	.4026076	.2567312	1.57	0.117	-.1005764	.9057916
govt	1.177792	.5421253	2.17	0.030	.1152461	2.240338
capital1	-.0281145	.0572111	-0.49	0.623	-.1402462	.0840173
_cons	14.63026	10.26693	1.42	0.154	-5.492552	34.75306

```
Endogenous variables:  consump wagepriv
Exogenous variables:   wagegovt govt capital1
```

Figura 31: Regresiones de mínimos cuadrados de tres estados ajustado en Stata

El comando de Stata **reg3** estima los coeficientes del modelo propuesto anteriormente.

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/klein.dta")

library(systemfit)

eq1 <- consump ~ wagepriv + wagegovt
eq2 <- wagepriv ~ consump + govt + capital1
inst <- ~wagegovt + govt + capital1
system <- list( eq1 = eq1, eq2 = eq2 )
fit3sls <- systemfit(system, "3SLS", inst = inst, data = datos)
summary(fit3sls)
```

```
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept) 19.355900   3.856336 5.01925 7.6157e-05 ***
## wagepriv     0.801275   0.137663 5.82056 1.3168e-05 ***
## wagegovt     1.029531   0.328027 3.13855  0.0054092 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.911394 on 19 degrees of freedom
## Number of observations: 22 Degrees of Freedom: 19
## SSR: 69.415117 MSE: 3.653427 Root MSE: 1.911394
## Multiple R-Squared: 0.938775 Adjusted R-Squared: 0.932331
##
```

```
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 14.7997848 11.3505151  1.30389 0.208702
## consump      0.4033571  0.2838271  1.42114 0.172371
## govt         1.1784058  0.5993421  1.96617 0.064896 .
## capital1    -0.0291787  0.0632493 -0.46133 0.650089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.620995 on 18 degrees of freedom
## Number of observations: 22 Degrees of Freedom: 18
## SSR: 123.653075 MSE: 6.869615 Root MSE: 2.620995
## Multiple R-Squared: 0.854439 Adjusted R-Squared: 0.830179
```

Figura 32: Regresiones de mínimos cuadrados de tres estados ajustado en R

En R en cambio se utilizó el comando **`systemfit`** el cual se basa en el método de mínimos cuadrados. De los resultados se observa que tanto en R como en Stata se obtienen los mismos valores en los coeficientes y sus parámetros.

Modelos de regresión lineal con efectos de tratamiento endógeno (etregress[25])

Descripción

Este modelo estima los parámetros de una regresión con efectos promedio en el tratamiento aumentado con una variable binaria endógena. Esta variable endógena debe estar correlacionada con el tratamiento pero no con el error ni las variables explicativas del modelo principal.

En este ejemplo se utilizará un subconjunto de la base de datos sobre ingresos de mujeres de Estados Unidos en 1972 con edades entre 18 y 30 años, para modelar los efectos promedios en el tratamiento para la variable **union** (pertenece o no a un sindicato) sobre el ingreso. Las variables a ser incluidas en el modelo son las siguientes:

Variable	**Descripción**	**Tipo de variable**
wage	Ingreso	Numérica
grade	Años de instrucción	Numérica
smsa	Variable indicativa de pertenencia a un distrito estadístico	Binaria
black	Variable indicativa para Afro-Americanos	Binaria
tenure	Permanencia en el trabajo actual	Numérica
south	Variable indicativa para residencia en el sur	Binaria

De estas variables se utilizará south, black y tenure, para modelar la variable endógena unión.

Modelo

El modelo efectos de tratamiento endógeno está compuesto de una ecuación para el resultado y_j y una ecuación para el tratamiento endógeno t_j,

$$y_j = \mathbf{x_j}\boldsymbol{\beta} + \delta t_j + \epsilon_j$$

$$t_j = \begin{cases} 1 & \text{si } \mathbf{w_j}\boldsymbol{\gamma} + u_j > 0 \\ 0 & \text{en otro caso} \end{cases}$$

[25] http://www.stata.com/manuals13/teetregress.pdf

donde $\mathbf{x_j}$ son las covariables del modelo principal, $\mathbf{w_j}$ son covariables usadas para modelar el tratamiento, y los términos de error ϵ_j y u_j tienen una distribución normal bivariada con media cero y matriz de covarianza

$$\begin{bmatrix} \sigma^2 & \rho\sigma = \lambda \\ \rho\sigma = \lambda & 1 \end{bmatrix}$$

Las covariables $\mathbf{x_j}$ y $\mathbf{w_j}$ son exógenas, es decir que no están correlacionadas con el término de error. En el ejemplo posterior, esto sería:

$$\text{wage} = -4.35 + 0.148\text{age} + 0.420\text{grade} + 0.911\text{smsa}$$

$$-0.788\text{black} + 0.152\text{tenure} + 2.94\text{union}$$

$$\text{union} = \begin{cases} 1 & \text{si } -.885 - 0.580\text{south} + 0.455\text{black} + 0.087\text{tenure} > 0 \\ 0 & \text{en otro caso} \end{cases}$$

$$\begin{bmatrix} 2.01^2 & -.574 * 2.01 = -1.160 \\ -.574 * 2.01 = -1.160 & 1 \end{bmatrix}$$

Se observa que los coeficientes son siginificaticos y que el efecto de tratamiento endógeno (coeficiente de union) es 2.95

Aplicación

Para estimar el modelo se empleará el comando **`etregress`** en Stata y el comando **`etreg`** del paquete del mismo nombre en R.

En la figura 33 y 34 se muestran los resultados tanto en R como en Stata de esta estimación:

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/union3.dta")
```

```
library(etreg)
fit1 = etreg(wage~age+grade+smsa+black+tenure,union~south+black+tenure,
             data=datos)
summary(fit1)
```

```
## Coefficients:
##              Estimate Std. error t value  Pr(> t)
## (Intercept) -4.35157    0.55433  -7.850 4.15e-15 ***
## age          0.14874    0.01980   7.511 5.88e-14 ***
## grade        0.42057    0.02983  14.097  < 2e-16 ***
## smsa         0.91170    0.12512   7.287 3.18e-13 ***
## black       -0.78825    0.13686  -5.759 8.44e-09 ***
## tenure       0.15240    0.03690   4.130 3.63e-05 ***
## union        2.94581    0.27650  10.654  < 2e-16 ***
## (Intercept) -0.88558    0.07260 -12.198  < 2e-16 ***
## south       -0.58074    0.08523  -6.814 9.48e-12 ***
## black        0.45575    0.09571   4.762 1.92e-06 ***
## tenure       0.08715    0.02327   3.745 0.000181 ***
## athRho      -0.65443    0.09136  -7.163 7.89e-13 ***
## lnsigma      0.70268    0.02945  23.862  < 2e-16 ***
## rho         -0.57465    0.06119  -9.391  < 2e-16 ***
## sigma        2.01915    0.05946  33.959  < 2e-16 ***
## lambda      -1.16030    0.15012  -7.729 1.08e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Newton-Raphson maximisation, 11 iterations
## Return code 2: successive function values within tolerance limit
## Log-likelihood: -3051.575 on 13 Df
```

Figura 33: Regresión lineal con efectos de tratamiento endógeno ajustado en R

```
. use http://www.stata-press.com/data/r13/union3
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)
. etregress wage age grade smsa black tenure, treat(union = south black tenure)
Iteration 0:   log likelihood = -3097.9871
Iteration 1:   log likelihood = -3052.5988
Iteration 2:   log likelihood = -3051.5789
Iteration 3:   log likelihood =  -3051.575
Iteration 4:   log likelihood =  -3051.575
Linear regression with endogenous treatment      Number of obs   =       1210
Estimator: maximum likelihood                    Wald chi2(6)    =     681.89
Log likelihood =  -3051.575                      Prob > chi2     =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
wage						
age	.1487409	.0193291	7.70	0.000	.1108566	.1866252
grade	.4205658	.0293577	14.33	0.000	.3630258	.4781058
smsa	.9117045	.1249041	7.30	0.000	.6668969	1.156512
black	-.7882471	.1367078	-5.77	0.000	-1.056189	-.5203047
tenure	.1524015	.0369596	4.12	0.000	.0799621	.2248409
union	2.945815	.2749624	10.71	0.000	2.406898	3.484731
_cons	-4.351572	.5283952	-8.24	0.000	-5.387208	-3.315936
union						
south	-.5807419	.0851111	-6.82	0.000	-.7475567	-.4139271
black	.4557499	.0958042	4.76	0.000	.2679772	.6435226
tenure	.0871536	.0232483	3.75	0.000	.0415878	.1327195
_cons	-.8855759	.0724506	-12.22	0.000	-1.027576	-.7435754
/athrho	-.6544344	.0910315	-7.19	0.000	-.8328529	-.4760159
/lnsigma	.7026768	.0293372	23.95	0.000	.645177	.7601767
rho	-.5746476	.0609711			-.6820049	-.4430472
sigma	2.01915	.0592362			1.906324	2.138654
lambda	-1.1603	.1495099			-1.453334	-.867266

```
LR test of indep. eqns. (rho = 0):   chi2(1) =    19.84   Prob > chi2 = 0.0000
```

Figura 34: Regresión lineal con efectos de tratamiento endógeno ajustado en Stata

Todas las variables del modelo son estadísticamente significativas y el valor de los valores de los parámetros coincide.

MODELOS DE SELECCIÓN Y NO PARAMÉTRICO

Este grupo de modelos permiten que el investigador pueda corregir el sesgo de selección (error en la elección de los individuos o grupos a participar en un estudio científico). El último de los siguientes modelos corresponde a un ajuste no parámetrico llamado regresión cuantíl.

Modelo de selección de Heckman (heckman[26])

Descripción

Incluye modelos de regresión donde la variable dependiente no es siempre observada, usando un estimador consistente de Heckman.

En el ejemplo se utilizará los datos del ingreso de 2000 Mujeres en Estados Unidos en 1972. Las variables seleccionadas son:

Variable	**Descripción**	**Tipo de variable**
wage	Ingreso	Numérica
education	Nivel de instrucción	Numérica
married	Variable indicativa de si está casada	Binaria
children	Número de hijos	Numérica

Se asumirá que el ingreso es una función de la educación y la edad, mientras que la probabilidad de trabajar es una función del estado marital, el número de hijos e implícitamente el ingreso (vía la inclusión de edad y educación, que se piensa determinan el ingreso).

Modelo

El modelo de selección introducido por Heckman (1976) asume que existe una relación de la forma:

$$y_j = \mathbf{x_j}\boldsymbol{\beta} + u_{1j} \quad \text{ecuación de regresión}$$

donde la variable dependiente y_j no se observa para todos los individuos. Sin embargo, es observada para la observación j si:

$$\mathbf{z_j}\boldsymbol{\gamma} + u_{2j} \quad \text{ecuación de selección}$$

[26] http://www.stata.com/manuals13/rheckman.pdf

donde $u_1 \sim N(0, \sigma)$, $u_2 \sim N(0,1)$ y $corr(u_1, u_2) = \rho$. Es decir que si ρ es significativo en el modelo, la estimación por éste método es adecuada. En el ejemplo se tiene:

$$\text{wage} = 0.485 + 0.989\text{educ} + 0.213\text{age}, \quad \text{ecuación de regresión}$$

$$-2.49 + 0.445\text{married} + 0.438\text{children} + 0.055\text{educ} + 0.036\text{age} \quad \text{ecuación de selección}$$

Se puede apreciar que $\rho = 0.703$ y es estadísticamente significativo por lo que se debe usar éste método para la estimación.

El modelo ajustado para el siguiente ejemplo es:

$$\text{wage} = \beta_0 + \beta_1\text{educ} + \beta_2\text{age} + u_1$$

Donde se asume que el ingreso fue observado si:

$$\gamma_0 + \gamma_1\text{married} + \gamma_2\text{children} + \gamma_3\text{educ} + \gamma_4\text{age} + u_2 > 0$$

Donde μ_1 y μ_2 tienen correlación ρ

Aplicación

Para estimar el modelo se utilizará el comando **Heckman** de Stata. En R se utilizará el comando **selection** del paquete **sampleSelection**. En la figura 35 y 36 se muestran los resultados obtenidos.

```
datos$selection = -as.integer(is.na(datos$wage)) + 1
fit1 = selection(selection ~ married + children + education + age,
                 wage ~ education+age, data = datos)
summary(fit1)
```

```
## --------------------------------------------
## Tobit 2 model (sample selection model)
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 4 iterations
## Return code 1: gradient close to zero
## Log-Likelihood: -5178.304
## 2000 observations (657 censored and 1343 observed)
## 10 free parameters (df = 1990)
## Probit selection equation:
##              Estimate Std. error t value  Pr(> t)
## (Intercept) -2.491015   0.189340 -13.156  < 2e-16 ***
## married      0.445171   0.067395   6.605 3.97e-11 ***
## children     0.438707   0.027783  15.791  < 2e-16 ***
## education    0.055732   0.010735   5.192 2.08e-07 ***
## age          0.036510   0.004153   8.790  < 2e-16 ***
## Outcome equation:
##             Estimate Std. error t value Pr(> t)
## (Intercept)  0.48579    1.07704   0.451   0.652
## education    0.98995    0.05326  18.588  <2e-16 ***
## age          0.21313    0.02060  10.345  <2e-16 ***
## Error terms:
##       Estimate Std. error t value Pr(> t)
## sigma  6.00479    0.16572   36.23  <2e-16 ***
## rho    0.70350    0.05123   13.73  <2e-16 ***
```

Figura 35: Modelo de selección de Heckman ajustado en R

```
. heckman wage educ age, select(married children educ age)
Iteration 0:   log likelihood = -5178.7009
Iteration 1:   log likelihood = -5178.3049
Iteration 2:   log likelihood = -5178.3045
Heckman selection model                         Number of obs     =      2000
(regression model with sample selection)        Censored obs      =       657
                                                Uncensored obs    =      1343
                                                Wald chi2(2)      =    508.44
Log likelihood = -5178.304                      Prob > chi2       =    0.0000
```

wage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
wage						
education	.9899537	.0532565	18.59	0.000	.8855729	1.094334
age	.2131294	.0206031	10.34	0.000	.1727481	.2535108
_cons	.4857752	1.077037	0.45	0.652	-1.625179	2.59673
select						
married	.4451721	.0673954	6.61	0.000	.3130794	.5772647
children	.4387068	.0277828	15.79	0.000	.3842534	.4931601
education	.0557318	.0107349	5.19	0.000	.0346917	.0767718
age	.0365098	.0041533	8.79	0.000	.0283694	.0446502
_cons	-2.491015	.1893402	-13.16	0.000	-2.862115	-2.119915
/athrho	.8742086	.1014225	8.62	0.000	.6754241	1.072993
/lnsigma	1.792559	.027598	64.95	0.000	1.738468	1.84665
rho	.7035061	.0512264			.5885365	.7905862
sigma	6.004797	.1657202			5.68862	6.338548
lambda	4.224412	.3992265			3.441942	5.006881

LR test of indep. eqns. (rho = 0): chi2(1) = 61.20 Prob > chi2 = 0.0000

Figura 36: Modelo de selección de Heckman ajustado en Stata

De lo que se observa los parámetros del modelo son estadísticamente significativos. Se puede observar que los valores tanto de los parámetros como de los test de las variables numéricas de las figuras 35 y 36 coinciden.

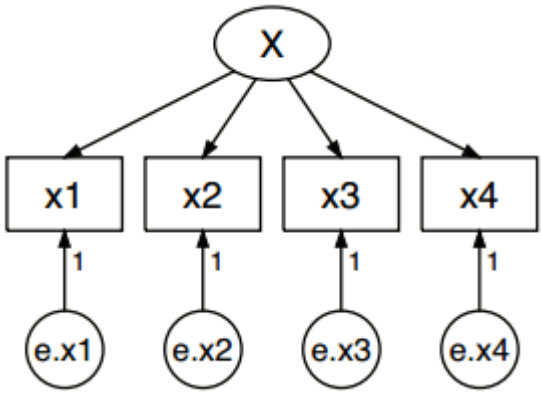
Modelos lineales y generalizados de ecuaciones estructurales (sem y gsem[27])

Descripción

Los modelos de ecuaciones estructurales incluyen un amplio conjunto de modelos desde regresiones lineales a modelos de medida y ecuaciones simultaneas, incluyendo análisis de factores confirmatorio, modelos de correlación única, modelos de crecimiento latente, indicadores múltiples y modelos de causas múltiples (MIMIC); y modelos basados en teoría de elemento-respuesta (IRT).

Modelo

En el siguiente ejemplo se ajustará un modelo de un factor de medida, según el siguiente esquema:



Modelo de un factor de Medida. Este modelo corresponde al siguiente conjunto de ecuaciones:

[27] http://www.stata.com/manuals13/semintro1.pdf#semintro1

$$x_1 = \alpha_1 + X\beta_1 + e.x_1$$

$$x_2 = \alpha_2 + X\beta_2 + e.x_2$$

$$x_3 = \alpha_3 + X\beta_3 + e.x_3$$

$$x_4 = \alpha_4 + X\beta_4 + e.x_4$$

El modelo se ajustará sobre una variable X ficticia perteneciente a la base de datos sem_1fmm.dta, parte de las bases de datos estándar de la ayuda de Stata 13.

Aplicación

Stata provee de dos comandos para el ajuste de este tipo de modelos **`gsem y sem`**. En R en cambio se utilizó el comando **`cfa`** del paquete lavaan (aún en desarrollo se encuentra en etapa de pruebas). En las figuras 37 y 38 se muestran los resultados de la estimación del modelo.

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/sem_1fmm.dta")
head(datos)
```

```
library(lavaan)
```

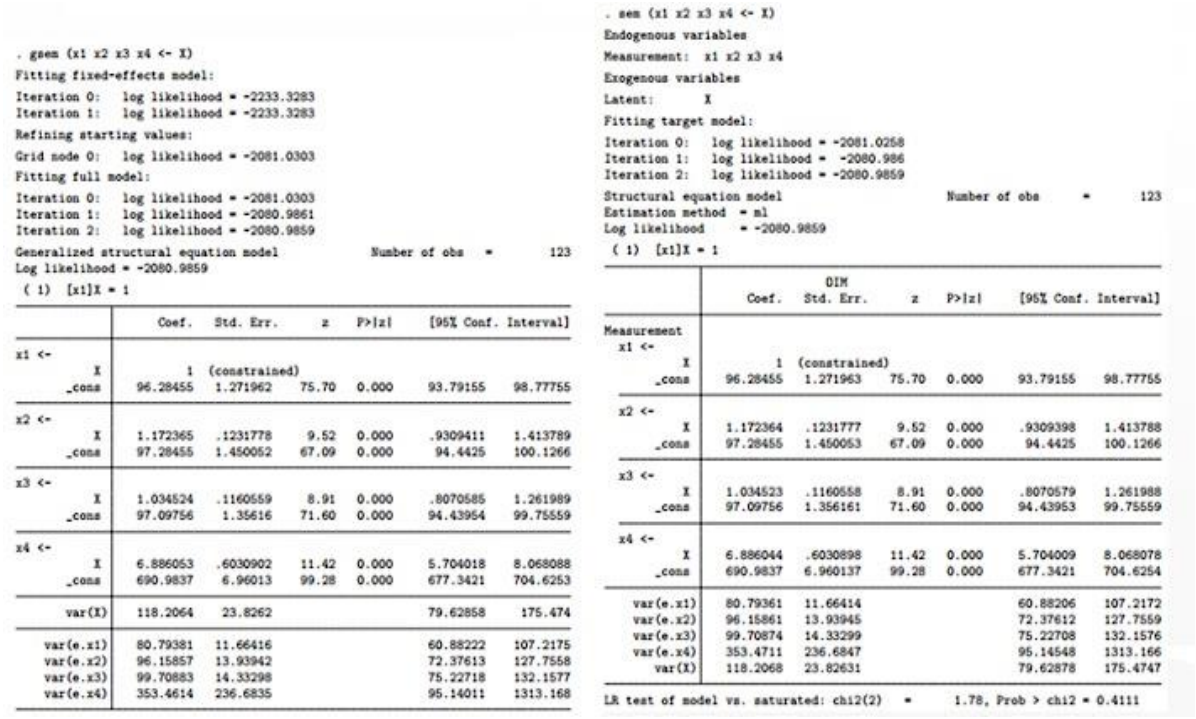
```
## This is lavaan 0.5-18
## lavaan is BETA software! Please report any bugs.
```

```
HS.model = "X =~ x1
            X =~ x2
            X =~ x3
            X =~ x4"
fit = cfa(model = HS.model, data = datos, estimator = "MLM")
summary(fit)
```

```
## lavaan (0.5-18) converged normally after  57 iterations
##
##   Number of observations                           123
##
##   Estimator                                         ML      Robust
##   Minimum Function Test Statistic                1.778       1.648
##   Degrees of freedom                                 2           2
##   P-value (Chi-square)                           0.411       0.439
##   Scaling correction factor                                  1.079
##     for the Satorra-Bentler correction
```

```
## Parameter estimates:
##
##   Information                                 Expected
##   Standard Errors                           Robust.sem
##
##                    Estimate  Std.err  Z-value  P(>|z|)
## Latent variables:
##   X =~
##     x1                1.000
##     x2                1.172    0.112   10.466    0.000
##     x3                1.035    0.118    8.759    0.000
##     x4                6.886    0.525   13.124    0.000
##
## Intercepts:
##     x1               96.285    1.277   75.389    0.000
##     x2               97.285    1.456   66.817    0.000
##     x3               97.098    1.362   71.306    0.000
##     x4              690.984    6.989   98.873    0.000
##     X                 0.000
##
## Variances:
##     x1               80.794   11.817
##     x2               96.159   12.210
##     x3               99.709   12.648
##     x4              353.472  222.685
##     X               118.207   22.164
```

Figura 37: Modelos generalizados de ecuaciones estructurales ajustado en R



```
. gsem (x1 x2 x3 x4 <- X)
Fitting fixed-effects model:
Iteration 0:   log likelihood = -2233.3283
Iteration 1:   log likelihood = -2233.3283
Refining starting values:
Grid node 0:   log likelihood = -2081.0303
Fitting full model:
Iteration 0:   log likelihood = -2081.0303
Iteration 1:   log likelihood = -2080.9861
Iteration 2:   log likelihood = -2080.9859
Generalized structural equation model           Number of obs   =        123
Log likelihood = -2080.9859
 ( 1)  [x1]X = 1
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x1 <-						
X	1	(constrained)				
_cons	96.28455	1.271962	75.70	0.000	93.79155	98.77755
x2 <-						
X	1.172365	.1231778	9.52	0.000	.9309411	1.413789
_cons	97.28455	1.450052	67.09	0.000	94.4425	100.1266
x3 <-						
X	1.034524	.1160559	8.91	0.000	.8070585	1.261989
_cons	97.09756	1.35616	71.60	0.000	94.43954	99.75559
x4 <-						
X	6.886053	.6030902	11.42	0.000	5.704018	8.068088
_cons	690.9837	6.96013	99.28	0.000	677.3421	704.6253
var(X)	118.2064	23.8262			79.62858	175.474
var(e.x1)	80.79381	11.66416			60.88222	107.2175
var(e.x2)	96.15857	13.93942			72.37613	127.7558
var(e.x3)	99.70883	14.33298			75.22718	132.1577
var(e.x4)	353.4614	236.6835			95.14011	1313.168

```
. sem (x1 x2 x3 x4 <- X)
Endogenous variables
Measurement:  x1 x2 x3 x4
Exogenous variables
Latent:       X
Fitting target model:
Iteration 0:   log likelihood = -2081.0258
Iteration 1:   log likelihood =  -2080.986
Iteration 2:   log likelihood = -2080.9859
Structural equation model                       Number of obs   =        123
Estimation method  = ml
Log likelihood     = -2080.9859
 ( 1)  [x1]X = 1
```

	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Measurement						
x1 <-						
X	1	(constrained)				
_cons	96.28455	1.271963	75.70	0.000	93.79155	98.77755
x2 <-						
X	1.172364	.1231777	9.52	0.000	.9309398	1.413788
_cons	97.28455	1.450053	67.09	0.000	94.4425	100.1266
x3 <-						
X	1.034523	.1160558	8.91	0.000	.8070579	1.261988
_cons	97.09756	1.356161	71.60	0.000	94.43953	99.75559
x4 <-						
X	6.886044	.6030898	11.42	0.000	5.704009	8.068078
_cons	690.9837	6.960137	99.28	0.000	677.3421	704.6254
var(e.x1)	80.79361	11.66414			60.88206	107.2172
var(e.x2)	96.15861	13.93945			72.37612	127.7559
var(e.x3)	99.70874	14.33299			75.22708	132.1576
var(e.x4)	353.4711	236.6847			95.14548	1313.166
var(X)	118.2068	23.82631			79.62878	175.4747

```
LR test of model vs. saturated: chi2(2)   =     1.78, Prob > chi2 = 0.4111
```

Figura 38: Modelos generalizados de ecuaciones estructurales ajustado en Stata

Los parámetros del modelo son estadísticamente significativos. Se puede observar que los valores tanto de los betas como de los test de las figuras 35 y 36 coinciden.

Modelos de regresión cuantílica (qreg[28])

Descripción

El objetivo es estimar modelos de regresión cuantílica que expresan los cuantiles de la distribución condicional como funciones lineales de las variables independientes. Por ejemplo, un objetivo particular es estimar la mediana de la variable dependiente en función de las dependientes.

Los datos usados para ilustrar el método corresponden a un conjunto de datos presentados en Cameron and Trivedi (2010, chap. 7). Las variables son x y y donde se tienen 5 observaciones para cada grupo.

[28] http://www.stata.com/manuals13/rqreg.pdf

Modelo

Puesto de manera sencilla, si se realiza una regresión sobre la mediana (o cualquier cuantil) se encuentra una línea que se ajusta a los datos de modo que se minimiza el balor absoluto de los residuos. Lo cual es la diferencia con la regresión *tradicional* donde se minimiza los residuos al cuadrado.

En el ejemplo se tiene:

$$y_{\text{mediana}} = 3 + 17x$$

x toma los valores 0 y 1. Si $x = 0$, la mediana del grupo 0 es igual a 3. Si $x = 1$, la mediana del grupo es $17 + 3 = 20$.

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. En las figuras 39 y 40 a continuación se muestran los resultados:

```
. qreg y x
Iteration  1:  WLS sum of weighted deviations =  60.941342

Iteration  1: sum of abs. weighted deviations =       55.5
Iteration  2: sum of abs. weighted deviations =         55

Median regression                               Number of obs =         10
  Raw sum of deviations     78.5 (about 14)
  Min sum of deviations       55                Pseudo R2     =     0.2994

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |          17   18.23213     0.93   0.378    -25.04338    59.04338
       _cons |           3   12.89207     0.23   0.822    -26.72916    32.72916
------------------------------------------------------------------------------
```

Figura 39: Regresión cuantílica ajustado en Stata

Los modelos de regresión cuantifica ajustados por **`qreg`**. Por defecto se estima una regresión con respecto a la mediana.

```
fit1 <- rq(y ~ x, tau = .5, data = datos)
summary(fit1)

## Warning in rq.fit.br(x, y, tau = tau, ci = TRUE, ...): Solution may be
## nonunique

##
## Call: rq(formula = y ~ x, tau = 0.5, data = datos)
##
## tau: [1] 0.5

##
## Coefficients:
##             coefficients lower bd upper bd
## (Intercept)  3.00000      0.42096 56.69245
## x           17.00000     15.02990 20.94020
```

Figura 40: Regresión cuantílica ajustado en R

En cambio en R se lo adaptó con el comando **`rq`** el cual se basa en el método de mínimos cuadrados. De los resultados se observa que hay una pequeña diferencia entre R y Stata ya que los métodos de optimización son de minimización de las desviaciones absolutas ponderadas y mínimos cuadrados respectivamente.

SERIES DE TIEMPO

Este grupo de modelos responde a series de tiempo. Mismas que son una colección de variables aleatorias indexadas en el tiempo. La mayoría de las veces el propósito es la predicción.

Modelos de Series de Tiempo con volatilidad variable (arch[29])

Descripción

El objetivo en este tipo de modelos es obtener una estimación de la volatilidad (varianza) de la serie de tiempo analizada. En la literatura se lo aborda como un modelo llamado GARCH donde la varianza condicional del cuadrado de las perturbaciones (ARCH) y de las varianzas condicionales en períodos anteriores.

[29] http://www.stata.com/manuals13/tsarch.pdf

En particular, para ejemplificar el uso de estos modelos, se dispone de datos del Índice de Precios al Por Mayor de Estados Unidos (Wholesale Price Index, WPI); Los datos se presentan de forma trimestral desde el primer trimestre de 1960 al cuarto trimestre de 1990. El modelo se ajustará sobre la tasa de cambio compuesta continua en el WPI: $\ln(\text{wpi}_t) - \ln(\text{wpi}_{t-1})$:

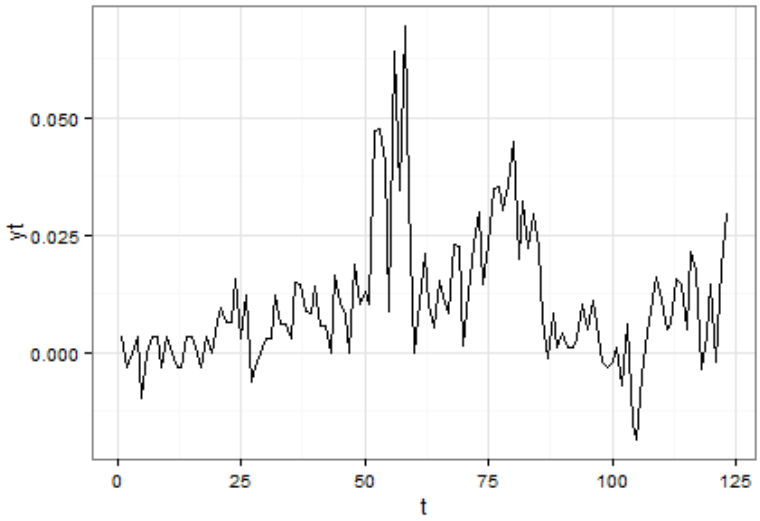


Figura 41: Serie de tiempo de la Tasa de cambio compuesta continua del WPI

Modelo

Se desea modelar la variabilidad de la serie. Como se aprecia en la figura 41esta serie contiene períodos de alta volatilidad seguidos de períodos de tranquilidad. El modelo a utilizar será un modelo ARCH de primer orden para la varianza condicional, que equivale al modelo GARCH(1,1).

El modelo ARCH, generalizado al GARCH(m, k), ajusta modelos de heterocedasticidad condicional autoregresiva:

$$y_t = \mathbf{x_t}\beta + \epsilon_t$$

$$Var(\varepsilon_t) = \sigma_t^2 = \gamma_0 + \gamma_1\epsilon_{t-1}^2 + \cdots + \gamma_m\epsilon_{t-m}^2 + \delta_1\sigma_{t-1}^2 + \cdots + \delta_k\sigma_{t-k}^2$$

donde γ_i son los parámetros de la parte ARCH y δ_i son los parámtros de la parte GARCH.

La ecuación para y_t puede contener opcionalmente términos ARMA (autoregresivos y de promedios móviles).

Los resultados obtenidos muestran un modelo estacionario, con parámetros iguales a:

$$y_t = 0.0061 + \epsilon_t$$

$$Var(\epsilon_t) = \sigma_t^2 = 0.436\epsilon_{t-1}^2 + 0.454\sigma_{t-1}^2$$

$$\text{donde } y_t = \ln(wpi_t) - \ln(wpi_{t-1})$$

Aplicación

Este modelo se puede ajustar en Stata mediante el comando **`arch`**, que tiene como equivalente en R, el paquete **`ugarch`**. Los resultados del ajuste se presentan en la figura 42 y 43.

```
spec1 <- ugarchspec(variance.model = list(garchOrder=c(1,1)),
                    mean.model = list(armaOrder=c(0,0), include.mean=TRUE))
fit1 <- ugarchfit(spec1, y)
fit1
```

```
## Optimal Parameters
## ------------------------------------
##         Estimate  Std. Error  t value Pr(>|t|)
## mu      0.006092    0.001255   4.8541 0.000001
## omega   0.000027    0.000012   2.1922 0.028368
## alpha1  0.445343    0.150118   2.9666 0.003011
## beta1   0.454492    0.114362   3.9742 0.000071
##
## Robust Standard Errors:
##         Estimate  Std. Error  t value Pr(>|t|)
## mu      0.006092    0.002613   2.3311 0.019747
## omega   0.000027    0.000020   1.3458 0.178377
## alpha1  0.445343    0.113455   3.9253 0.000087
## beta1   0.454492    0.108974   4.1706 0.000030
##
## LogLikelihood : 373.2433
```

Figura 42: Series de Tiempo con volatilidad variable ajustado en R

```
. arch D.ln_wpi, arch(1) garch(1)
(setting optimization to BHHH)
Iteration 0:   log likelihood =  355.23458
Iteration 1:   log likelihood =  365.64586
  (output omitted)
Iteration 10:  log likelihood =  373.23397
ARCH family regression
Sample: 1960q2 - 1990q4                         Number of obs   =       123
Distribution: Gaussian                          Wald chi2(.)    =         .
Log likelihood =   373.234                      Prob > chi2     =         .
```

D.ln_wpi	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
ln_wpi						
_cons	.0061167	.0010616	5.76	0.000	.0040361	.0081974
ARCH						
arch L1.	.4364123	.2437428	1.79	0.073	-.0413147	.9141394
garch L1.	.4544606	.1866606	2.43	0.015	.0886127	.8203086
_cons	.0000269	.0000122	2.20	0.028	2.97e-06	.0000508

Figura 43: Series de Tiempo con volatilidad variable ajustado en Stata

De lo que se observa los parámetros del modelo son estadísticamente significativos.

Modelos de regresión para series de tiempo, con términos autoregresivos y de promedios móviles (arima[30])

Descripción

El acrónimo ARIMA se deriva de "Auto-Regressive Integrated Moving Average" o modelos autoregresivos integrados y de promedios móviles. Los rezagos de la serie estacionarizada en la ecuación de pronóstico se conocen como términos autoregresivos y los rezagos de los errores de pronóstico se denominan términos de promedios móviles.

Continuando con el ejemplo anterior, se emplearán los datos sobre el Índice de Precios al por mayor de los Estados Unidos (WPI). Se calculará primeras diferencias sobre la serie para

[30] http://www.stata.com/manuals13/tsarima.pdf

convertirla en estacionaria, junto con un término de rezago y promedios móviles, resultando en un modelo ARIMA(1, 1, 1).

Modelo

Un modelo arima no estacional o ARIMA (p, d, q), está dado por la siguiente ecuación:

$$y_t = \alpha + \rho_1 y_{t-1} + \cdots + \rho_p y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

Donde la serie y_t corresponde al cálculo de d veces las diferencias de la serie original, de tal modo que esta sea estacionaria.

Aplicación

En las figuras 44 y 45 se muestra un ejemplo que considera un modelo ARIMA (1,1,1)

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/wpi1.dta")
```

```
fit1 = armaFit( formula=~arima(1,1,1), data=datos$wpi)
summary(fit1)
```

```
## Moments:
## Skewness Kurtosis
##  0.07495  1.29021
##
## Coefficient(s):
##      Estimate  Std. Error  t value Pr(>|t|)
## ar1   0.94122     0.03787   24.853  < 2e-16 ***
## ma1  -0.46574     0.10462   -4.452 8.52e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## sigma^2 estimated as: 0.5398
## log likelihood:       -137.25
## AIC Criterion:        280.49
```

Figura 44: Regresión para series de tiempo, con términos autoregresivos y de promedios móviles ajustado en R

En R el comando utilizado para estimar el modelo se encuentra en el paquete **`fArma`** que añade algunas opciones útiles a la función **`arima`** del paquete **`Stats`**.

```
. use http://www.stata-press.com/data/r13/wpi1
. arima wpi, arima(1,1,1)
(setting optimization to BHHH)
Iteration 0:   log likelihood = -139.80133
Iteration 1:   log likelihood =  -135.6278
Iteration 2:   log likelihood = -135.41838
Iteration 3:   log likelihood = -135.36691
Iteration 4:   log likelihood = -135.35892
(switching optimization to BFGS)
Iteration 5:   log likelihood = -135.35471
Iteration 6:   log likelihood = -135.35135
Iteration 7:   log likelihood = -135.35132
Iteration 8:   log likelihood = -135.35131
ARIMA regression
Sample:  1960q2 - 1990q4                        Number of obs      =       123
                                                Wald chi2(2)       =    310.64
Log likelihood = -135.3513                      Prob > chi2        =    0.0000
```

D.wpi	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
wpi						
_cons	.7498197	.3340968	2.24	0.025	.0950019	1.404637
ARMA						
ar L1.	.8742288	.0545435	16.03	0.000	.7673256	.981132
ma L1.	-.4120458	.1000284	-4.12	0.000	-.6080979	-.2159938
/sigma	.7250436	.0368065	19.70	0.000	.6529042	.7971829

```
Note: The test of the variance against zero is one sided, and the two-sided
    confidence interval is truncated at zero.
```

Figura 45: Regresión para series de tiempo, con términos autoregresivos y de promedios móviles ajustado en Stata

Es Stata el comando utilizado es **`arima`**. La diferencia que se aprecia en los coeficientes estimados se debe a que R no estima un intercepto para series diferenciadas. Se aprecia en el resultado que ambos coeficientes son significativos.

El modelo se puede contrastar contra otros modelos ARIMA utilizando los índices AIC y el logaritmo de máxima verosimilitud.

Modelos de regresión para errores estándar de Newey-West (newey)

Descripción

El cálculo de errores robustos por el método de Newey-West se recomienda en regresiones donde la estructura del error es heterocedástica y posiblemente correlacionada con algún rezago.

Para el ejemplo se utilizará la base de datos idle.dta, misma que describe la distribución de tiempo de un procesador entre tiempo inactivo y tiempo de usuario. En este modelo se considerará en el cálculo del error una correlación hasta el tercer rezago.

Aplicación

Para estimar los errores de regresión robustos se utilizará el comando **`newey`** de Stata y el comando **`NeweyWest`** de R. En las figuras 46 y 47se muestran los resultados del ejemplo.

```
library("sandwich")
sqrt(diag(NeweyWest(reg1, lag=3, prewhite = FALSE)))
```

```
## (Intercept)        idle
##  6.11249302  0.06674994
```

Figura 46: Regresión para errores estándar de Newey-West ajustado en R

```
. use http://www.stata-press.com/data/r13/idle2, clear
. tsset time
        time variable:  time, 1 to 30
                delta:  1 unit
. newey usr idle, lag(3)
Regression with Newey-West standard errors              Number of obs    =        30
maximum lag: 3                                          F(  1,    28)    =     10.90
                                                        Prob > F         =    0.0026
------------------------------------------------------------------------------
             |             Newey-West
         usr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        idle |  -.2281501   .0690927    -3.30   0.003    -.3696801    -.08662
       _cons |   23.13483   6.327031     3.66   0.001     10.17449    36.09516
------------------------------------------------------------------------------
```

Figura 47: Regresión para errores estándar de Newey-West ajustado en Stata

El comando NeweyWest de R devuelve la matriz de varianza covarianza, razón por la cual es necesario calcular la raíz cuadrada de la diagonal para calcular los errores estándar de las estimaciones. Además de ve que los parámetros son significativos y coinciden

DATOS DE PANEL

El siguiente grupo de modelos se usan para conjunto de individuos medidos en el tiempo. En la literatura son más conocidos como datos de panel.

Modelos lineales de efectos aleatorios (xtreg[31])

Descripción

Se quiere ajustar un modelo de regresión de datos de panel con efectos aleatorios. Usualmente los efectos aleatorios están asociados a la variable que identifica el panel. Desde luego, también se pueden estimar los efectos fijos del modelo.

Para el ejemplo se usa los datos de una encuesta longitudinal nacional de 1968 en mujeres de 14 a 26 años[32] en relación a su situación laboral. Las variables para ajustar el modelo son:

Variable	Descripción	Tipo de variable
ln_wage	Log natural de (salario/deflactor del PNB[33])	Numérica
age	Edad	Numérica
grade	Grado completo (años de escolaridad completa, entre 1 y 18)	Numérica
ttl_exp	Experiencia total de trabajo	
tenure	Antigüedad	Factor (Binaria)
race	Etnia (1 = blanco, 2 = afro, 3= otra)	Factor (3 niveles)
not_smsa	Personas que viven fuera de un área estadística metropolitana estándar.	Factor (Binaria)
south	Si es del Sur	Factor (Binaria)

Debe aclararse que en modelo se usan interacciones. `c.age#c.age` es la edad al cuadrado, `c.ttl_exp#c.ttl_exp` es la experiencia a cuadrado, `c.tenure#c.tenure` es la antigüedad al cuadrado. Se usa esa notación particular para aclarar que se las use como factores. Se desea explicar el logaritmo del salario (el porcentaje del salario) que se ve afectado por las demás variables. Es decir, el salario en términos de años completos de escolaridad (grado), edad actual y la edad al cuadrado; años actuales trabajadas (experiencia) y la experiencia al cuadrado; años actuales de la tenencia en el trabajo actual y la tenencia al

[31] http://www.stata.com/manuals13/xtxtreg.pdf
[32] Los datos se pueden encontrar en: http://www.bls.gov/nls/nlsorig.htm
[33] Producto Nacional Bruto.

cuadrado. La variable de panel se llama `idcode`, la cual representa el código de la encuesta (4148 encuestas).

Modelo

Considere modelos de la forma

$$y_{it} = \alpha + \mathbf{x_{it}}\boldsymbol{\beta} + \nu_i + \epsilon_{it}$$

En este modelo, $\nu_i + \epsilon_{it}$ es el término de error (en el cual no estamos interesados). Es de interés estimar $\boldsymbol{\beta}$. ν_i es el error por unidad/individuo; difiere para diferentes unidades pero para cualquier unidad en particular, su valor es constante.

Según los resultados de las estimaciones del ejemplo, se tiene:

$$\begin{gathered}\ln_w = 0.3339 + 0.0607\text{grade} + 0.0323\text{age} - .0005\text{age}^2 \\ +0.0138\text{ttl}_e\text{xp} + 0.0007\text{ttl}_e\text{xp} + 0.06984\text{tenure} - 0.0028\text{tenure}^2 \\ -0.0564\text{black} - 0.186\text{not}_s\text{msa} - 0.099\text{south}\end{gathered}$$

Se aprecia que el modelo es significativo ($F = 450.2$).

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. Las figuras 48 y 49 a continuación se muestran los resultados:

```
. xtreg ln_w grade age c.age#c.age ttl_exp c.ttl_exp#c.ttl_exp tenure
> c.tenure#c.tenure 2.race not_smsa south, be
Between regression (regression on group means)  Number of obs      =     28091
Group variable: idcode                          Number of groups   =      4697
R-sq:  within  = 0.1591                         Obs per group: min =         1
       between = 0.4900                                        avg =       6.0
       overall = 0.3695                                        max =        15
                                                F(10,4686)         =    450.23
sd(u_i + avg(e_i.))=  .3036114                  Prob > F           =    0.0000
```

ln_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
grade	.0607602	.0020006	30.37	0.000	.0568382	.0646822
age	.0323158	.0087251	3.70	0.000	.0152105	.0494211
c.age#c.age	-.0005997	.0001429	-4.20	0.000	-.0008799	-.0003194
ttl_exp	.0138853	.0056749	2.45	0.014	.0027598	.0250108
c.ttl_exp# c.ttl_exp	.0007342	.0003267	2.25	0.025	.0000936	.0013747
tenure	.0698419	.0060729	11.50	0.000	.0579361	.0817476
c.tenure# c.tenure	-.0028756	.0004098	-7.02	0.000	-.0036789	-.0020722
race						
black	-.0564167	.0105131	-5.37	0.000	-.0770272	-.0358061
not_smsa	-.1860406	.0112495	-16.54	0.000	-.2080949	-.1639862
south	-.0993378	.010136	-9.80	0.000	-.1192091	-.0794665
_cons	.3339113	.1210434	2.76	0.006	.0966093	.5712133

Figura 48: Modelos lineales de efectos aleatorios ajustado en Stata

El comando de Stata **`xtreg`** ajusta modelos de regresión a datos de panel. en particular xtreg con la opción **be** ajusta modelos de efectos aleatorios usando el estimador de regresión *between*, con la opción **fe** ajusta modelos de efectos fijos[34]. Se reportan los valores de los coeficientes bajo el supuesto *between*, así como su valor de R cuadrado. Por otro lado, en R se tiene:

[34] Si se deseara ajustar un modelo de efectos aleatorios, por ejemplo al asumir efectos aleatorios la opción en lugar de **`be`**, sería **`re theta`**.

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/nlswork.dta")

library(plm)

## Loading required package: Formula

datos$raceB = as.factor(datos$race == 2)
fit1 = plm(ln_wage ~ grade + age + I(age^2) + ttl_exp + I(ttl_exp^2) +
             tenure + I(tenure^2) + not_smsa + south + raceB,
           model = "between", data = datos)
summary(fit1)
```

```
## Coefficients :
##                 Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept)   0.33391131  0.12104342   2.7586 0.0058273 **
## grade         0.06076019  0.00200056  30.3716 < 2.2e-16 ***
## age           0.03231577  0.00872510   3.7038 0.0002149 ***
## I(age^2)     -0.00059966  0.00014294  -4.1952 2.777e-05 ***
## ttl_exp       0.01388530  0.00567493   2.4468 0.0144505 *
## I(ttl_exp^2)  0.00073417  0.00032673   2.2471 0.0246829 *
## tenure        0.06984188  0.00607291  11.5006 < 2.2e-16 ***
## I(tenure^2)  -0.00287556  0.00040978  -7.0174 2.584e-12 ***
## not_smsa     -0.18604056  0.01124954 -16.5376 < 2.2e-16 ***
## south        -0.09933778  0.01013600  -9.8005 < 2.2e-16 ***
## raceBTRUE    -0.05641669  0.01051306  -5.3663 8.422e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    846.98
## Residual Sum of Squares: 431.95
## R-Squared      :  0.49
##       Adj. R-Squared :  0.48886
## F-statistic: 450.23 on 10 and 4686 DF, p-value: < 2.22e-16
```

Figura 49: Modelos lineales de efectos aleatorios ajustado en R

En R en cambio se utilizó el comando **`plm`** del paquete con el mismo nombre el cual se basa en el método de los mínimos cuadrados para su implementación. De lo que se observa que se obtienen los mismos resultados en sus coeficientes tanto en R como en Stata.

Modelo de regresión lineal dinámica de Arellano–Bond para datos de panel (xtabond[35])

Descripción

El objetivo es ajustar modelos datos de panel dinámicos lineales incluyen *t*_rezagos de la variable dependiente como covariables y contienen efectos no observados a nivel de panel, fijos o aleatorios. Por construcción los objetos no observados a nivel de panel están correlacionados con las variables dependientes rezagadas, haciendo que los estimadores estándar sean inconsistentes

Los datos usados para ejemplificar el método corresponden a un panel desbalanceado de empresas de Inglaterra. Las variables son:

Variable	Descripción	Tipo de variable
n	Ln del empleo de la empresa	Numérica
w	Ln del salario	Numérica

[35] http://www.stata.com/manuals13/xtxtabond.pdf

k	Ln del stock de capital bruto	Numérica
ys	Ln de la producción de la empresa	Numérica

El modelo también incluye variables binarias para los años `yr1980, yr1981, yr1982, yr1983, yr1984`. Se desea conocer el empleo de la empresa en función de las demás variables de la tabla anterior.

Modelo

Un modelo de datos de panel dinámico tiene la forma:

$$y_{it} = \sum_{j=1}^{p} \alpha_j \, y_{i,t-j} + \mathbf{x_{it}}\boldsymbol{\beta_1} + \mathbf{w_{it}}\boldsymbol{\beta_2} + \nu_i + \epsilon_{it}$$

donde $i = 1, \dots, N$ y $t = 1, \dots, T_i$; α_j son los parámetros a ser estimados, $\mathbf{x_{it}}$ es un vector $1 \times k_1$ de covariables estrictamente exógenas; $\boldsymbol{\beta_1}$ es un vector $k_1 \times 1$ de parámetros a ser estimados; $\mathbf{w_{it}}$ es un vector $1 \times k_2$ de covariables endógenas predeterminado; $\boldsymbol{\beta_2}$ es un vector $k_2 \times 1$ de parámetros a ser estimados; ν_i son los efectos del panel (que podrían estár correlacionados con las covariables); y ϵ_{it} son independientes e idénticamente distribuidos sobre toda la muestra con varianza σ_ϵ^2. Los ν_i y ϵ_{it} se asumen independientes para cada i y sobre todo t.

Usando los resultados del ajuste se tiene:

$$n_{it}$$
$$= 0.708 n_{i,t-1} - 0.0886 n_{i,t-2} - 0.605 + 0.4096 w_{i,t-1} - 0.355 - 0.0599 k_{i,t-1}$$
$$- 0.0211 k_{i,t-2} + 0.6264 - 0.7231 ys_{i,t-1} + 0.1179 ys_{i,t-2} + 0.0113 yr1980$$
$$- 0.0212 yr1981 - 0.03495 yr1982 - 0.0287 yr1983 - 0.0148 yr1984$$

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. En las figuras 50 y 51 a continuación se muestran los resultados:

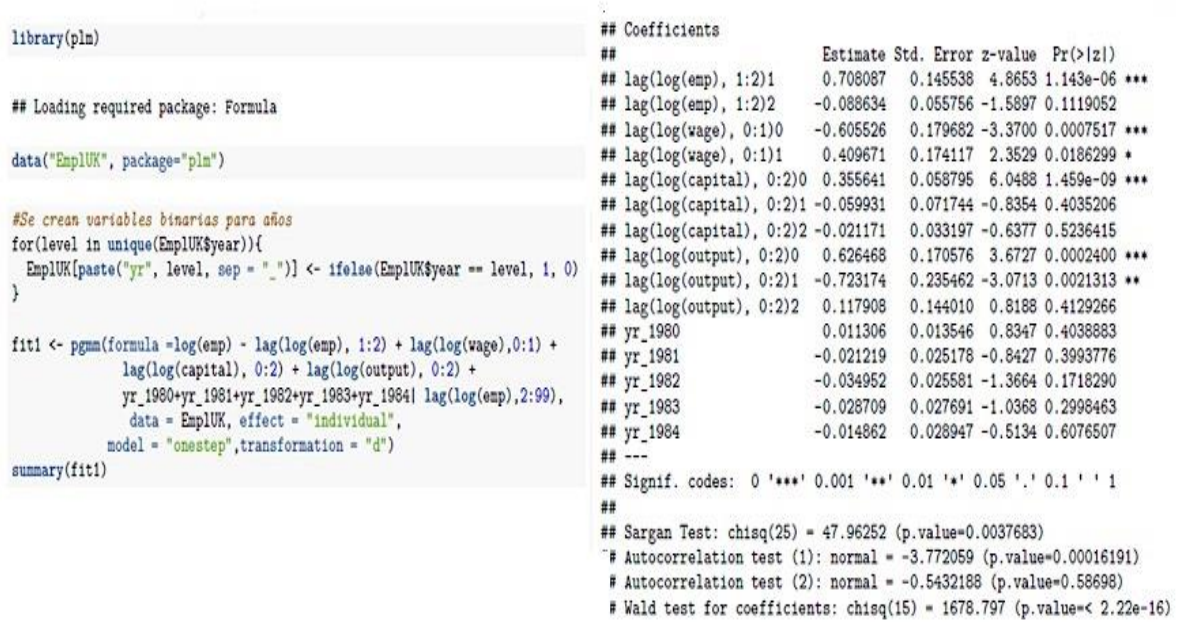
```
. xtabond n l(0/1).w l(0/2).(k ys) yr1980-yr1984, lags(2) noconstant

Arellano-Bond dynamic panel-data estimation  Number of obs      =       611
Group variable: id                           Number of groups   =       140
Time variable: year
                                             Obs per group:  min =         4
                                                             avg =  4.364286
                                                             max =         6

Number of instruments =     40               Wald chi2(15)      =   1627.13
                                             Prob > chi2        =    0.0000
One-step results
```

n	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
n						
L1.	.7080866	.1455545	4.86	0.000	.4228051	.9933681
L2.	-.0886343	.0448479	-1.98	0.048	-.1765346	-.000734
w						
--.	-.605526	.0661129	-9.16	0.000	-.735105	-.4759471
L1.	.4096717	.1081258	3.79	0.000	.1977491	.6215943
k						
--.	.3556407	.0373536	9.52	0.000	.2824289	.4288525
L1.	-.0599314	.0565918	-1.06	0.290	-.1708493	.0509865
L2.	-.0211709	.0417927	-0.51	0.612	-.1030831	.0607412
ys						
--.	.6264699	.1348009	4.65	0.000	.3622651	.8906748
L1.	-.7231751	.1844696	-3.92	0.000	-1.084729	-.3616214
L2.	.1179079	.1440154	0.82	0.413	-.1643572	.400173
yr1980	.0113066	.0140625	0.80	0.421	-.0162554	.0388686
yr1981	-.0212183	.0206559	-1.03	0.304	-.0617031	.0192665
yr1982	-.034952	.022122	-1.58	0.114	-.0783103	.0084063
yr1983	-.0287094	.0251536	-1.14	0.254	-.0780096	.0205909
yr1984	-.014862	.0284594	-0.52	0.602	-.0706414	.0409174

```
Instruments for differenced equation
        GMM-type: L(2/.).n
        Standard: D.w LD.w D.k LD.k L2D.k D.ys LD.ys L2D.ys D.yr1980
                  D.yr1981 D.yr1982 D.yr1983 D.yr1984
```

Figura 50: Regresión lineal dinámica de Arellano–Bond para datos de panel ajustado en Stata

En comando de Stata **`xtabond`** se declaran un rezagos para el empleo (`lags(2)`), un rezago para el salario y dos rezagos para el stock de capital y la producción. Finalmente, se usan algunos años como variables binarias (de 1980 a 1984). La variable de panel es el identificador de la empresa `id` y la variable de tiempo es `year`. En `R` se tiene:



```
library(plm)

## Loading required package: Formula

data("EmplUK", package="plm")

#Se crean variables binarias para años
for(level in unique(EmplUK$year)){
  EmplUK[paste("yr", level, sep = "_")] <- ifelse(EmplUK$year == level, 1, 0)
}

fit1 <- pgmm(formula =log(emp) ~ lag(log(emp), 1:2) + lag(log(wage),0:1) +
               lag(log(capital), 0:2) + lag(log(output), 0:2) +
               yr_1980+yr_1981+yr_1982+yr_1983+yr_1984| lag(log(emp),2:99),
              data = EmplUK, effect = "individual",
            model = "onestep",transformation = "d")
summary(fit1)
```

```
## Coefficients
##                       Estimate Std. Error z-value  Pr(>|z|)
## lag(log(emp), 1:2)1    0.708087   0.145538  4.8653 1.143e-06 ***
## lag(log(emp), 1:2)2   -0.088634   0.055756 -1.5897 0.1119052
## lag(log(wage), 0:1)0  -0.605526   0.179682 -3.3700 0.0007517 ***
## lag(log(wage), 0:1)1   0.409671   0.174117  2.3529 0.0186299 *
## lag(log(capital), 0:2)0 0.355641  0.058795  6.0488 1.459e-09 ***
## lag(log(capital), 0:2)1 -0.059931 0.071744 -0.8354 0.4035206
## lag(log(capital), 0:2)2 -0.021171 0.033197 -0.6377 0.5236415
## lag(log(output), 0:2)0  0.626468  0.170576  3.6727 0.0002400 ***
## lag(log(output), 0:2)1  -0.723174 0.235462 -3.0713 0.0021313 **
## lag(log(output), 0:2)2  0.117908  0.144010  0.8188 0.4129266
## yr_1980                 0.011306  0.013546  0.8347 0.4038883
## yr_1981                -0.021219  0.025178 -0.8427 0.3993776
## yr_1982                -0.034952  0.025581 -1.3664 0.1718290
## yr_1983                -0.028709  0.027691 -1.0368 0.2998463
## yr_1984                -0.014862  0.028947 -0.5134 0.6076507
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Sargan Test: chisq(25) = 47.96252 (p.value=0.0037683)
# Autocorrelation test (1): normal = -3.772059 (p.value=0.00016191)
 # Autocorrelation test (2): normal = -0.5432188 (p.value=0.58698)
 # Wald test for coefficients: chisq(15) = 1678.797 (p.value=< 2.22e-16)
```

Figura 51: Regresión lineal dinámica de Arellano–Bond para datos de panel ajustado en R

En R se utilizó el comando **pgmm** para el ajuste el cual se basa en el método de mínimos cuadrados. De lo que se observa los valores de los test de Wald es apenas distitno.

Modelo de estimación en datos dinámicos de panel (xtdpd[36])

Descripción

El objetivo es ajustar modelos de panel dinámicos lineales que incluyen *t*_rezagos tanto de la variable dependiente como de las covariables. Además contienen efectos no observados a nivel de panel, fijos o aleatorios.

Los datos usados para ejemplificar el método corresponden a un panel desbalanceado de empresas de Inglaterra. Las variables son:

Variable	Descripción	Tipo de variable
N	Ln del empleo de la empresa	Numérica
W	Ln del salario	Numérica
K	Ln del stock de capital bruto	Numérica
ys	Ln de la producción de la empresa	Numérica

[36] http://www.stata.com/manuals13/xtxtdpd.pdf

El modelo también incluye variables binarias para los años `yr1980, yr1981, yr1982, yr1983, yr1984`. Se desea conocer el empleo de la empresa en función de las demás variables de la tabla anterior.

Modelo

Considere un modelo de panel dinámico tiene la forma:

$$y_{it} = \sum_{j=1}^{p} \alpha_j \, y_{i,t-j} + \mathbf{x_{it}}\boldsymbol{\beta_1} + \mathbf{w_{it}}\boldsymbol{\beta_2} + \nu_i + \epsilon_{it}$$

donde $i = 1, \dots, N$ y $t = 1, \dots, T_i$; α_j son los parámetros a ser estimados, $\mathbf{x_{it}}$ es un vector $1 \times k_1$ de covariables estrictamente exógenas; $\boldsymbol{\beta_1}$ es un vector $k_1 \times 1$ de parámetros a ser estimados; $\mathbf{w_{it}}$ es un vector $1 \times k_2$ de covariables endógenas predeterminado; $\boldsymbol{\beta_2}$ es un vector $k_2 \times 1$ de parámetros a ser estimados; ν_i son los efectos del panel (que podrían estár correlacionados con las covariables); y ϵ_{it} son iid o proceder de un proceso de medias móviles con varianza σ_ϵ^2.

Usando los resultados del ajuste se tiene:

$$\begin{gathered} n_{it} = 0.625 n_{i,t-1} - 0.0809 n_{i,t-2} \\ -0.7734 + 0.204 w_{i,t-1} \\ +0.614 - 0.564 ys_{i,t-1} \\ +0.369 + 0.006 k_{i,t-1} - 0.020 k_{i,t-2} \\ +0.0128 yr1980 - 0.0018 yr1981 + 0.0038 yr1982 + 0.0289 yr1983 + 0.0481 yr1984 \end{gathered}$$

Las variables estrictamente endógenas se declaran con div().

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. En las figuras 52 y 53 a continuación se muestran los resultados:

```
. xtdpd L(0/2).n L(0/1).(w ys) L(0/2).k yr1980-yr1984, div(L(0/1).(ys) yr1980-yr1984) dgmmiv(n)
> step noconstant vce(robust)

Dynamic panel-data estimation                 Number of obs         =       611
Group variable: id                            Number of groups      =       140
Time variable: year
                                              Obs per group:    min =         4
                                                                avg =  4.364286
                                                                max =         6

Number of instruments =    100                Wald chi2(14)         =    763.91
                                              Prob > chi2           =    0.0000
Two-step results
                                    (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |              WC-Robust
           n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
         L1. |   .6250383   .1217039     5.14   0.000      .386503    .8635735
         L2. |  -.0809401   .0729133    -1.11   0.267    -.2238475    .0619673
             |
           w |
         --. |  -.7734339   .1009608    -7.66   0.000    -.9713134   -.5755545
         L1. |   .2046473    .113593     1.80   0.072     -.017991    .4272855
             |
          ys |
         --. |   .6147036   .1639069     3.75   0.000     .2934519    .9359553
         L1. |  -.5643852   .1634929    -3.45   0.001    -.8848253   -.2439451
             |
           k |
         --. |    .369188   .0979728     3.77   0.000      .177165    .5612111
         L1. |   .0068721   .0745421     0.09   0.927    -.1392277     .152972
         L2. |  -.0208353   .0418027    -0.50   0.618     -.102767    .0610964
             |
      yr1980 |    .012818   .0138382     0.93   0.354    -.0143043    .0399403
      yr1981 |  -.0018814   .0256565    -0.07   0.942    -.0521672    .0484043
      yr1982 |   .0038399   .0283345     0.14   0.892    -.0516946    .0593744
      yr1983 |   .0289356    .033138     0.87   0.383    -.0360137    .0938849
      yr1984 |    .048176   .0332323     1.45   0.147    -.0169582    .1133101
------------------------------------------------------------------------------
Instruments for differenced equation
        GMM-type: L(2/.).n L(1/.).w L(1/.).k
        Standard: D.ys LD.ys D.yr1980 D.yr1981 D.yr1982 D.yr1983 D.yr1984
```

Figura 52: Estimación en datos dinámicos de panel ajustado en Stata

Note que en el comando **`xtdpd`** se declaran dos rezagos para el empleo (variable que se desea explicar), un rezago para el salario y la producción y dos rezagos para el stock de capital. Finalmente, se usan algunos años como variables binarias (de 1980 a 1984). La variable de panel es el identificador de la empresa `id` y la variable de tiempo es `year`. Finalmente se menciona que este comando es muy parecido a `xtabond`, pero es un tanto más general ya que permite medias móviles en los errores del modelo. En `R` se tiene:

```
fit2 <- pgmm(formula =log(emp) ~ lag(log(emp), 1:2) + lag(log(wage),0:1) +
             lag(log(capital), 0:2) + lag(log(output), 0:1) +
       yr1980+yr1981+yr1982+yr1983+yr1984|
       lag(log(emp),2:99)+lag(log(wage),1:99)+lag(log(capital),1:99)|
       log(output)+lag(log(output))+yr1980+yr1981+yr1982+yr1983+
       yr1984,data = EmplUK, effect = "individual",
     model = "twostep",transformation = "d")
summary(fit2)
## Coefficients
##                              Estimate Std. Error z-value  Pr(>|z|)
## lag(log(emp), 1:2)1         0.6250386  0.1217039  5.1357 2.810e-07 ***
## lag(log(emp), 1:2)2        -0.0809401  0.0729133 -1.1101 0.2669612
## lag(log(wage), 0:1)0       -0.7734340  0.1009608 -7.6607 1.849e-14 ***
## lag(log(wage), 0:1)1        0.2046470  0.1135929  1.8016 0.0716112
## lag(log(capital), 0:2)0     0.3691878  0.0979728  3.7683 0.0001644 ***
## lag(log(capital), 0:2)1     0.0068722  0.0745421  0.0922 0.9265456
## lag(log(capital), 0:2)2    -0.0208354  0.0418025 -0.4984 0.6181847
## lag(log(output), 0:1)0      0.6147030  0.1639068  3.7503 0.0001766 ***
## lag(log(output), 0:1)1     -0.5643842  0.1634926 -3.4520 0.0005564 ***
## yr1980                      0.0128179  0.0138382  0.9263 0.3543038
## yr1981                     -0.0018815  0.0256565 -0.0733 0.9415396
## yr1982                      0.0038399  0.0283345  0.1355 0.8922006
## yr1983                      0.0289356  0.0331380  0.8732 0.3825619
## yr1984                      0.0481760  0.0332323  1.4497 0.1471491
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Sargan Test: chisq(86) = 91.8632 (p.value=0.31282)
## Autocorrelation test (1): normal = -2.721464 (p.value=0.0064994)
## Autocorrelation test (2): normal = -0.425446 (p.value=0.67051)
## Wald test for coefficients: chisq(14) = 763.9142 (p.value=< 2.22e-16)
```

Figura 53: Estimación en datos dinámicos de panel ajustado en R

El comando `pgmm` en R fue utilizado para el ajuste. De lo que se observa que se obtienen los mismos resultados en sus coeficientes y sus tests.

Modelos de frontera estocástica de datos de panel (xtfrontier[37])

Descripción

El objetivo es ajustar modelos de producción o coste estocásticos para datos de panel. más específicamente, se estima los parámetros de un modelo lineal con perturbaciones (errores) generadas por distribuciones mixtas específicas. Esto significa que el error del modelo se asume con dos componentes: uno que proviene de una distribución no negativa (término de *ineficiencia*) y otro generado por una distribución simétrica (por esto se le llama *mixto*).

Los datos usados para ilustrar el método corresponden a empresas (91 empresas) que tienen un producto (llamado *widget*). Las variables son:

Variable	Descripción	Tipo de variable

[37] http://www.stata.com/manuals13/xtxtfrontier.pdf

lnwidgets	Ln del número de productos	Numérica
lnmachines	Ln del número de horas por máquina usados en la producción	Numérica
lnworkers	Ln del número de horas-hombre requeridas para producir	Numérica

Es preciso mencionar que en este ejemplo de asume una especificación tiempo-invariante. Esto significa que se asume que el error del efecto aleatorio no cambia en el tiempo (término de ineficiencia).

Modelo

La estructura del modelo es igual a la usada en frontier pero sobre datos de panel. Así se tiene:

$$q_{it} = f(\mathbf{z_{it}}, \boldsymbol{\beta})\xi_{it}\exp(v_{it})$$

donde $\mathbf{z_{it}}$ son las covariables de las que depende la producción, $\boldsymbol{\beta}$ sus respectivos parámetros, ξ_{it} es el nivel de ineficiencia y $\exp(v_{it})$ son choques aleatorios. Note que si no existen choques ni ineficiencia, la producción sería óptima, es decir $q_{it} = f(\mathbf{z_{it}}, \boldsymbol{\beta})$.

Tomando logaritmos se tiene:

$$\ln(q_{it}) = \beta_0 + \sum_{j=1}^{k} \beta_j \ln(z_{jit}) + v_{it} - u_{it}$$

donde $u_{it} = -\ln(\xi_{it})$. Se asume que $v_i \sim N(0, \sigma_v)$ y que u_i puede tener distribución exponencial, media-normal o normla truncada. De manera similar se puede tener una formulación para el costo:

$$\ln(c_{it}) = \beta_0 + \beta_q \ln(q_{it}) \sum_{j=1}^{k} \beta_j \ln(p_{jit}) + v_{it} + u_{it}$$

En el ejemplo posterior se tiene las siguientes estimaciones:

$$\text{lnwidgets} = 3.030 + 0.290\text{lnmachines} + 0.2943\text{lnworkers}$$

Por defecto se asume que u_i tiene una distribución media-normal. Note que el estimador de la varianza del efecto de ineficiencia $\sigma_u^2 = 3.139$ es significativo, el estimador de la varianza del

choque aleatorio $\sigma_v^2 = 1.005$ también lo es. El parámetro $\mu = 1.125$ es la media global de la respuesta, también se reporta la estimación de la varianza $\sigma = 4.145$.

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. Las figuras 54 y 55 a continuación se muestran los resultados:

```
. use http://www.stata-press.com/data/r13/xtfrontier1
. xtfrontier lnwidgets lnmachines lnworkers, ti
Iteration 0:   log likelihood = -1473.8703
Iteration 1:   log likelihood = -1473.0565
Iteration 2:   log likelihood = -1472.6155
Iteration 3:   log likelihood =  -1472.607
Iteration 4:   log likelihood = -1472.6069
Time-invariant inefficiency model               Number of obs      =       948
Group variable: id                              Number of groups   =        91
                                                Obs per group: min =         6
                                                               avg =      10.4
                                                               max =        14
                                                Wald chi2(2)       =    661.76
Log likelihood  = -1472.6069                    Prob > chi2        =    0.0000
------------------------------------------------------------------------------
   lnwidgets |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  lnmachines |   .2904551   .0164219    17.69   0.000     .2582688    .3226415
   lnworkers |   .2943333   .0154352    19.07   0.000     .2640808    .3245858
       _cons |   3.030983   .1441022    21.03   0.000     2.748548    3.313418
-------------+----------------------------------------------------------------
         /mu |   1.125667   .6479217     1.74   0.082    -.144236    2.39557
   /lnsigma2 |   1.421979   .2672745     5.32   0.000      .898131    1.945828
  /ilgtgamma |   1.138685   .3562642     3.20   0.001     .4404204     1.83695
-------------+----------------------------------------------------------------
      sigma2 |   4.145318   1.107938                      2.455011    6.999424
       gamma |   .7574382   .0654548                      .6083592    .8625876
     sigma_u2 |   3.139822   1.107235                      .9696821    5.309962
     sigma_v2 |   1.005496   .0484143                      .9106055    1.100386
------------------------------------------------------------------------------
```

Figura 54: Modelos de frontera estocástica de datos de panel ajustado en Stata

El comando de Stata **`xtfrontier`** proporciona estimaciones de máxima verosimilitud de los parámetros del modelo de decaimiento invariante en el tiempo. Note que `sigma_v2` es la estimación de la varianza del error, `sigma_u2` es la estimación de la varianza del efecto aleatorio, `sigma2` es la suma de la varianza de `sigma_u2 + sigma_v2`, `gamma` es el cociente de las varianzas de `sigma_u2/sigma2`. En `R` tenemos:

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/xtfrontier1.dta")

library(plm)
library(frontier)
pData = plm.data(datos, "id")
fit1 <- sfa(lnwidgets~lnmachines+lnworkers, data = pData, truncNorm = TRUE)
summary(fit1)

## Error Components Frontier (see Battese & Coelli 1992)
## Inefficiency decreases the endogenous variable (as in a production function)
## The dependent variable is logged
## Iterative ML estimation terminated after 14 iterations:
## log likelihood values and parameters of two successive iterations
## are within the tolerance limit
##
## final maximum likelihood estimates
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept) 3.030953   0.125490 24.1529 < 2.2e-16 ***
## lnmachines  0.290455   0.016372 17.7407 < 2.2e-16 ***
## lnworkers   0.294333   0.015381 19.1363 < 2.2e-16 ***
## sigmaSq     4.145409   1.163257  3.5636 0.0003658 ***
## gamma       0.757445   0.066855 11.3296 < 2.2e-16 ***
## mu          1.125567   0.659594  1.7065 0.0879232 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## log likelihood value: -1472.607
##
## panel data
## number of cross-sections = 91
## number of time periods = 14
## total number of observations = 948
## thus there are 326 observations not in the panel
##
## mean efficiency: 0.2817233
```

Figura 55: Modelos de frontera estocástica de datos de panel ajustado en R

En R se hizo el ajuste con el comando **`sfa`** el cual se basa en el análisis de máxima verosimilitud. De lo que se observa que se obtienen los mismos resultados en sus coeficientes y tests.

Modelos de datos de panel para mínimos cuadrados generalizados (xtgls[38])

Descripción

El objetivo es ajustar ajusta modelos lineales de datos de panel usando mínimos cuadrados generalizados factibles. Se permite la estimación de los parámetros en presencia de autocorrelación dentro de los paneles y correlación cruzada y heterocedasticidad entre paneles.

Para ilustrar el método se usa un ejemplo en el que se trabaja con datos de un estudio clásico de la demanda de inversión por Grunfeld y Griliches (1960). Éstos corresponden a información de 10 empresas en más de 20 años, desde 1935 a 1954; en particular:

Variable	Descripción	Tipo de variable
Mvalue	Valor de mercado de la empresa	Numérica
Kstock	Valor de su stock de capital	Numérica

[38] http://www.stata.com/manuals13/xtxtgls.pdf

Invest	Inversión	Numérica

Se desea estimar la inversión de cada empresa en función de las demás variables de la tabla anterior permitiendo que exista correlación entre paneles.

Modelo

El modelo ajustado es:

$$\text{invest} = -38.361 + 0.0961\text{market} + 0.3095\text{stock}$$

En el ejemplo se ha especificado que se permite que exista heterocedasticidad y correlación entre los paneles.

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. Las figuras 56 y 57 a continuación se muestran los resultados:

```
. xtgls invest market stock, panels(correlated)
Cross-sectional time-series FGLS regression
Coefficients:  generalized least squares
Panels:        heteroskedastic with cross-sectional correlation
Correlation:   no autocorrelation
Estimated covariances      =        15          Number of obs      =       100
Estimated autocorrelations =         0          Number of groups   =         5
Estimated coefficients     =         3          Time periods       =        20
                                                Wald chi2(2)       =   1285.19
                                                Prob > chi2        =    0.0000
------------------------------------------------------------------------------
      invest |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      market |   .0961894   .0054752    17.57   0.000     .0854583    .1069206
       stock |   .3095321   .0179851    17.21   0.000     .2742819    .3447822
       _cons |  -38.36128   5.344871    -7.18   0.000    -48.83703   -27.88552
------------------------------------------------------------------------------
```

Figura 56: Datos de panel para mínimos cuadrados generalizados ajustado en Stata

Se precia que la opción `panels(correlated)` del comando **`xtgls`** es la que nos permite tener correlación entre los paneles. La variable de panel corresponde a la empresa y el tiempo es la variable `time`. En `R` se tiene:

```
library(rio)
Produc = import("http://www.stata-press.com/data/r13/invest2.dta")

library(panelAR)
Produc$time = as.integer(Produc$time)
fit1 <- panelAR(invest~stock+market, timeVar = "time",data = Produc, autoCorr='ar1',
                panelVar = "company", panelCorrMethod = "phet")
summary(fit1)

##
## Panel Regression with AR(1) Prais-Winsten correction and panel heteroskedasticity-robust standard
##
## Balanced Panel Design:
##  Total obs.:       100 Avg obs. per panel 20
##  Number of panels: 5   Max obs. per panel 20
##  Number of times:  20  Min obs. per panel 20
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -38.67030   39.32157  -0.983    0.328
## stock         0.35433    0.06012   5.894 5.46e-08 ***
## market        0.09310    0.01295   7.188 1.37e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-squared: 0.5793
## Wald statistic: 113.3905, Pr(>Chisq(2)): 0
```

Figura 57: Datos de panel para mínimos cuadrados generalizados ajustado en R

En R se lo realizó con **`panelAR`** el cual es un comando que se basa en el método de máxima verosimilitud. De lo que se ve en los resultados de R y Stata se ve una pequeña variación que depende del método de optimización de cada uno de los comandos y del hecho de que Stata usa el método de mínimos cuadrados generalizados para la estimación de parámetros.

Modelos de estimadores para errores de Hausman–Taylor (xthtaylor[39])

Descripción

El objetivo es ajustar modelos de datos de panel con efectos aleatorios en los cuales algunas de las covariables están correlacionadas con los efectos individuales aleatorios no observados pero no están correlacionadas con el error global del modelo.

Se tiene un ejemplo en el que se replica los resultados de Baltagi y Khanti-Akom, utilizando 595 observaciones a las personas mayores de 1976 a 1982 que fueron extraídos del Estudio de Panel de Dinámica de Ingresos (PSID). Las variables analizadas son

Variable	**Descripción**	**Tipo de variable**
ln_wage	Log natural de salario	Numérica
occ	Igual a 1 si es trabajador industrial	Factor (Binaria)
smsa	Igual a 1 si la persona vive en una zona metropolitana grande	Factor (Binaria)

[39] http://www.stata.com/manuals13/xtxthtaylor.pdf

south	1 si es del sur	Factor (Binaria)
ind	Igual a 1 si trabaja en el sector industrial	Factor (Binaria)
exp	Experiencia	Factor (3 niveles)
wks	Semanas trabajadas	Numérica
ms	Estado civil	Factor (Binaria)
union	Igual a 1 si el salario es fijado en un contrato colectivo	Factor (Binaria)
fem	Igual a 1 si es de género femenino	Factor (Binaria)
blk	Igual a 1 si es afroamericano	Factor (Binaria)
ed	Años de educación	Numérica

La idea central es que se asume que las variables `exp, exp2, wks, ms,` y `union` están correlacionadas con el efecto aleatorio producido por la variable `id` (identificador de persona). De desea ajustar un modelo que explique el logaritmo del salario en función de las demás variables.

Modelo

Considere el modelo de efectos aleatorios:

$$y_{it} = \mathbf{X_{1it}\beta_1} + \mathbf{X_{2it}\beta_2} + \mathbf{Z_{1i}\delta_1} + \mathbf{Z_{2i}\delta_2} + \mu_i + \epsilon_{it}$$

donde:

- $\mathbf{X_{1it}}$ es un vector $1 \times k_1$ de observaciones de variables exógenas, variantes en el tiempo que se asumen no correlacionadas con μ_i y ϵ_{it};
- $\mathbf{X_{2it}}$ es un vector $1 \times k_2$ de observaciones de variables endógenas, variantes en el tiempo que se asumen (posiblemente) correlacionadas con μ_i pero ortogonales a ϵ_{it};
- $\mathbf{Z_{1i}}$ es un vector $1 \times g_1$ de observaciones de variables exógenas, no variantes en el tiempo que se asumen no correlacionadas con μ_i y ϵ_{it};
- $\mathbf{Z_{2i}}$ es un vector $1 \times g_2$ de observaciones de variables endógenas, no variantes en el tiempo que se asumen (posiblemente) correlacionadas con μ_i pero ortogonales a ϵ_{it};
- μ_i es el efecto aleatorio no observado por panel que se asumen que tiene media cero y varianza finita σ_μ^2 y es i.i.d. en los paneles.

- ϵ_{it} es un error idiosincrático que se asume que tiene media cero y varianza finita σ_ϵ^2 y es i.i.d. sobre todas las observaciones.
- $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\delta_1}$ y $\boldsymbol{\delta_2}$ son vectores de coeficientes $k_1 \times 1, k_2 \times 1$, $g_1 \times 1$ y $g_2 \times 1$, respectivamente; y $i = 1, \ldots, n$, donde n es el número de paneles en la muestra y, para cada i, $t = 1, \ldots, T_i$.

Usando el modelo entonces se tendría:

$$\begin{gathered}\text{lwage} = 2.912 - 0.0207\text{occ} + 0.0074\text{south} - 0.0418\text{smsa} + 0.0136\text{ind} \\ +0.1131\text{exp} - 0.0004\text{exp2} + 0.0008\text{wks} - 0.0298\text{ms} + 0.0327\text{union} \\ -0.1309\text{fem} - 0.285\text{blk} \\ +0.137\text{ed}\end{gathered}$$

La fracción de la varianza debida al efecto de panel σ_μ^2 es $\rho = 0.9746$, lo que indica que se debe tomar en cuenta el efecto de panel para estos datos.

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. Las figuras 58 y 59 a continuación se muestran los resultados:

```
. xthtaylor lwage occ south smsa ind exp exp2 wks ms union fem blk ed,
> endog(exp exp2 wks ms union ed)
Hausman-Taylor estimation                       Number of obs      =      4165
Group variable: id                              Number of groups   =       595
                                                Obs per group: min =         7
                                                               avg =         7
                                                               max =         7
Random effects u_i ~ i.i.d.                     Wald chi2(12)      =   6891.87
                                                Prob > chi2        =    0.0000
```

lwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
TVexogenous						
occ	-.0207047	.0137809	-1.50	0.133	-.0477149	.0063055
south	.0074398	.031955	0.23	0.816	-.0551908	.0700705
smsa	-.0418334	.0189581	-2.21	0.027	-.0789906	-.0046761
ind	.0136039	.0152374	0.89	0.372	-.0162608	.0434686
TVendogenous						
exp	.1131328	.002471	45.79	0.000	.1082898	.1179758
exp2	-.0004189	.0000546	-7.67	0.000	-.0005259	-.0003119
wks	.0008374	.0005997	1.40	0.163	-.0003381	.0020129
ms	-.0298508	.01898	-1.57	0.116	-.0670508	.0073493
union	.0327714	.0149084	2.20	0.028	.0035514	.0619914
TIexogenous						
fem	-.1309236	.126659	-1.03	0.301	-.3791707	.1173234
blk	-.2857479	.1557019	-1.84	0.066	-.5909179	.0194221
TIendogenous						
ed	.137944	.0212485	6.49	0.000	.0962977	.1795902
_cons	2.912726	.2836522	10.27	0.000	2.356778	3.468674
sigma_u	.94180304					
sigma_e	.15180273					
rho	.97467788	(fraction of variance due to u_i)				

```
Note:  TV refers to time varying; TI refers to time invariant.
```

Figura 58: Estimadores para errores de Hausman–Taylor ajustado en Stata

Note que el porcentaje de la varianza que se le atribuye al efecto aleatorio es bastante alta (97%). Esto ayuda a fundamentar el uso de este modelo. Además, Stata clasifica las covariables en endógenas y exógenas. Recuerde que se declaró como endógenas a: `exp, exp2, wks, ms,` y `unión.`

```
library(rio)
wages =  import("http://www.stata-press.com/data/r13/psidextract.dta")

library(plm)
fit1 = pht(lwage~occ+south+smsa+ind+exp+exp2+wks+ms+union+fem+blk+ed|
           occ+south+smsa+ind+fem+blk, data=wages, index="id")
summary(fit1)
## Coefficients :
##                Estimate  Std. Error t-value  Pr(>|t|)
## (Intercept)  2.9127e+00  2.8365e-01 10.2687 < 2.2e-16 ***
## occ         -2.0705e-02  1.3781e-02 -1.5024   0.13299
## south        7.4398e-03  3.1955e-02  0.2328   0.81590
## smsa        -4.1833e-02  1.8958e-02 -2.2066   0.02734 *
## ind          1.3604e-02  1.5237e-02  0.8928   0.37196
## exp          1.1313e-01  2.4710e-03 45.7851 < 2.2e-16 ***
## exp2        -4.1886e-04  5.4598e-05 -7.6718 1.696e-14 ***
## wks          8.3740e-04  5.9973e-04  1.3963   0.16263
## ms          -2.9851e-02  1.8980e-02 -1.5728   0.11578
## union        3.2771e-02  1.4908e-02  2.1982   0.02794 *
## fem         -1.3092e-01  1.2666e-01 -1.0337   0.30129
## blk         -2.8575e-01  1.5570e-01 -1.8352   0.06647 .
## ed           1.3794e-01  2.1248e-02  6.4919 8.474e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    886.9
## Residual Sum of Squares: 95.947
## F-statistic: 2852.33 on 12 and 4152 DF, p-value: < 2.22e-16
```

Figura 59: Estimadores para errores de Hausman–Taylor ajustado en R

En R se utilizó el comando **`pht`** el que se basa en el método de máxima verosimilitud. De lo que se observa que se obtienen los mismos resultados en sus coeficientes como en sus parámetros.

Modelos de regresión de intervalos de datos de panel (xtintreg[40])

Descripción

El objetivo es ajustar modelos de regresión con efectos aleatorios cuyas variables dependientes pueden estar medidas como datos puntuales, datos de intervalo, datos censurados por la izquierda o datos censurados por la derecha para datos de panel. Esto significa:

Tipo de dato		$var.dep_{inferior}$	$var.dep_{superior}$
Puntual	$a = [a, a]$	a	a
De intervalo	$[a, b]$	a	b
Censurado por izquierda	$(-\infty, b]$	.	b
Censurado por derecha	$[a, -\infty)$	a	.

Para el ejemplo se usa los datos de una encuesta longitudinal nacional de 1968 en mujeres de 14 a 26 años[41]. Las variables para ajustar el modelo son:

Variable	**Descripción**	**Tipo de variable**
ln_wage	Log natural de (salario/deflactor del PNB[42])	Numérica
union	Pertenece a aluna asociación	Factor (Binaria)
age	Edad	Numérica
grade	Grado completo (años de escolaridad completa, entre 1 y 18)	Numérica
not_smsa	Personas que viven fuera de un área estadística metropolitana estándar.	Factor (Binaria)
south	1 si es del sur	Factor (Binaria)
year	Año de la entrevista, entre 1968 y 1988	Factor
occ_code	Código de ocupación	Factor (3 niveles)

[40] http://www.stata.com/manuals13/xtxtintreg.pdf
[41] Los datos se pueden encontrar en: http://www.bls.gov/nls/nlsorig.htm
[42] Producto Nacional Bruto.

Se desea explicar el logaritmo del salario (el porcentaje del salario) que se ve afectado por las demás variables. Cabe aclarar que en modelo también se usan las variables `south` y `year` como una iteracción (multiplicadas). La variable de panel se llama `idcode` genera 4140 grupos.

Modelo

Considere el modelo de regresión lineal con efectos aleatorios por panel:

$$y_{it} = \mathbf{x_{it}}\boldsymbol{\beta}\nu_i + \epsilon_{it}$$

para $i = 1, \dots, n$ paneles, donde $t = 1, \dots, n_i$. Los efectos aleatorios, ν_i, son i.i.d., $N(0, \sigma_\nu^2)$, y ϵ_{it} son i.i.d., $N(0, \sigma_\epsilon^2)$ independientes de ν_i. Los datos observados consisten de las parejas, (y_{1it}, y_{2it}) tal que todo lo que se sabe es que $(y_{1it} \leq y_{it} \leq y_{2it})$, donde y_{1it} podría ser $-\infty$ y y_{2it} podría ser $+\infty$.

Usando las estimaciones del ejemplo se tendría:

$$\ln_w \text{age} = 0.379 + 0.1441\text{union} + 0.0104\text{age} + 0.0794\text{grade}$$
$$-0.377\text{south} + 0.0013 + 0.0034\text{south} * \text{year} - 0.0197\text{occ}_\text{c}\text{ode}$$

La fracción de la varianza debida al efecto de panel $\sigma_\nu^2 = 0.298$ es $\rho = 0.417$ y significativa, lo que indica que se debe tomar en cuenta el efecto de panel para estos datos.

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. Las figuras 60 y 61 a continuación se muestran los resultados:

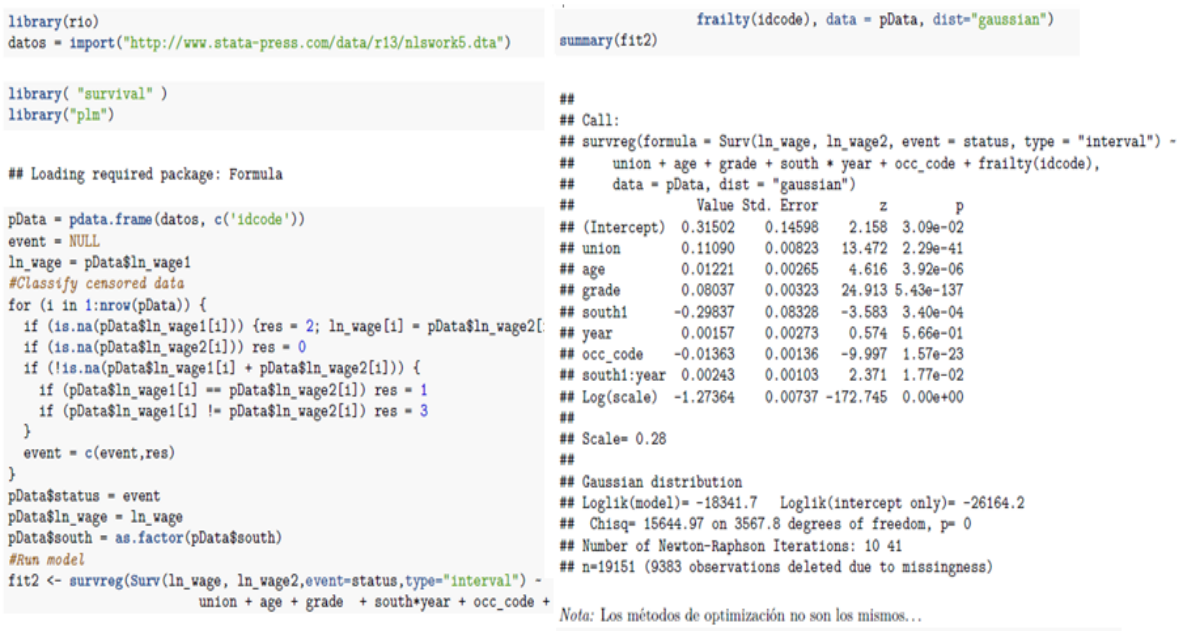
```
. use http://www.stata-press.com/data/r13/nlswork5
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)
. xtintreg ln_wage1 ln_wage2 union age grade south##c.year occ_code, intreg
  (output omitted)
Random-effects interval regression              Number of obs      =     19151
Group variable: idcode                          Number of groups   =      4140
Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       4.6
                                                               max =        12
Integration method: mvaghermite                 Integration points =        12
                                                Wald chi2(7)       =   2523.84
Log likelihood  = -23174.355                    Prob > chi2        =    0.0000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |   .1441844   .0094245    15.30   0.000     .1257128    .162656
         age |   .0104083   .0018804     5.54   0.000     .0067228    .0140939
       grade |   .0794958   .0023469    33.87   0.000      .074896    .0840955
    1.south  |  -.3778103   .0979415    -3.86   0.000    -.5697722   -.1858485
        year |   .0013528   .0020176     0.67   0.503    -.0026016    .0053071
             |
south#c.year |
          1  |   .0034385   .0012105     2.84   0.005     .0010659     .005811
             |
    occ_code |  -.0197912   .0014094   -14.04   0.000    -.0225535   -.0170289
       _cons |   .3791078   .1136641     3.34   0.001     .1563303    .6018853
-------------+----------------------------------------------------------------
     /sigma_u|   .2987074   .0052697    56.68   0.000     .2883789     .309036
     /sigma_e|   .3528109   .0030935   114.05   0.000     .3467478     .358874
-------------+----------------------------------------------------------------
         rho |   .4175266   .0102529                      .3975474    .4377211
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01)= 2516.85 Prob>=chibar2 = 0.000
```

Figura 60: Regresión de intervalos de datos de panel ajustado en Stata

Este modelo debe ser especificado de modo que la variable dependiente de cuenta el tipo de dato asociado a ella. Por ejemplo:

Tipo de dato	**ln_wage1**	**ln_wage2**
Puntual	1.451214	1.451214
Censurado por izquierda	1.02862	.
Censurado por derecha	.	1.589977
De Intervalo	1.7	1.8

Se precisa que el modelo considera efectos aleatorios para la variable de panel.



```
library(rio)
datos = import("http://www.stata-press.com/data/r13/nlswork5.dta")

library( "survival" )
library("plm")

## Loading required package: Formula

pData = pdata.frame(datos, c('idcode'))
event = NULL
ln_wage = pData$ln_wage1
#Classify censored data
for (i in 1:nrow(pData)) {
  if (is.na(pData$ln_wage1[i])) {res = 2; ln_wage[i] = pData$ln_wage2[
  if (is.na(pData$ln_wage2[i])) res = 0
  if (!is.na(pData$ln_wage1[i] + pData$ln_wage2[i])) {
    if (pData$ln_wage1[i] == pData$ln_wage2[i]) res = 1
    if (pData$ln_wage1[i] != pData$ln_wage2[i]) res = 3
  }
  event = c(event,res)
}
pData$status = event
pData$ln_wage = ln_wage
pData$south = as.factor(pData$south)
#Run model
fit2 <- survreg(Surv(ln_wage, ln_wage2,event=status,type="interval") ~
                      union + age + grade  + south*year + occ_code +
                      frailty(idcode), data = pData, dist="gaussian")
summary(fit2)

##
## Call:
## survreg(formula = Surv(ln_wage, ln_wage2, event = status, type = "interval") ~
##     union + age + grade + south * year + occ_code + frailty(idcode),
##     data = pData, dist = "gaussian")
##                Value Std. Error       z        p
## (Intercept)  0.31502    0.14598   2.158 3.09e-02
## union        0.11090    0.00823  13.472 2.29e-41
## age          0.01221    0.00265   4.616 3.92e-06
## grade        0.08037    0.00323  24.913 5.43e-137
## south1      -0.29837    0.08328  -3.583 3.40e-04
## year         0.00157    0.00273   0.574 5.66e-01
## occ_code    -0.01363    0.00136  -9.997 1.57e-23
## south1:year  0.00243    0.00103   2.371 1.77e-02
## Log(scale)  -1.27364    0.00737 -172.745 0.00e+00
##
## Scale= 0.28
##
## Gaussian distribution
## Loglik(model)= -18341.7   Loglik(intercept only)= -26164.2
##  Chisq= 15644.97 on 3567.8 degrees of freedom, p= 0
## Number of Newton-Raphson Iterations: 10 41
## n=19151 (9383 observations deleted due to missingness)
```

Nota: Los métodos de optimización no son los mismos…

Figura 61: Regresión de intervalos de datos de panel ajustado en R

En tanto que en R se utilizó el comando **`survreg`** el cual se basa en máxima verosimilitud.

De lo que se ve en los resultados de R y Stata se ve una pequeña variación que depende del método de optimización de cada uno de los comandos.

Modelos de regresión de variables instrumentales para datos de panel (xtivreg[43])

Descripción

El objetivo es ajustar un modelo de regresión de de datos de panel en los cuales algunas covariables del lado derecho son endógenas (instrumentales[44]), estos estimadores son generalizaciones del método de mínimos cuadrados en dos etapas para estimadores simples de datos de panel para variables exógenas. Desde luego, al ser un enfoque de datos de panel, se pueden estimar efectos fijos y aletarios.

[43] http://www.stata.com/manuals13/xtxtivreg.pdf
[44] Variable correlacionada con alguna covariable del modelo pero no correlacionada con el error del modelo principal.

Para el ejemplo se usa los datos de una encuesta longitudinal nacional de 1968 en mujeres de 14 a 26 años[45] en relación a su situación laboral. Las variables para ajustar el modelo son:

Variable	Descripción	Tipo de variable
ln_wage	Log natural de (salario/deflactor del PNB[46])	Numérica
age	Edad	Numérica
union	Pertenece a aluna asociación	Factor (Binaria)
tenure	Antigüedad	Factor (Binaria)
not_smsa	Personas que viven fuera de un área estadística metropolitana estándar.	Factor (Binaria)
south	Si es del Sur	Factor (Binaria)

En el código de Stata `c.age#c.age` es la edad al cuadrado. Se usa esa notación particular para aclarar que se las use como factores. Se desea explicar el logaritmo del salario (el porcentaje del salario) que se ve afectado por las demás variables. La variable de panel se llama `idcode`, la cual forma 4134 grupos.

Modelo

Considere el modelo de la forma:

$$y_{it} = \mathbf{Y_{it}\gamma} + \mathbf{X_{1it}\beta} + \mu_i + \nu_{it} = \mathbf{Z_{it}\delta} + \mu_i + \nu_{it}$$

donde:

- y_{it} es la variable dependiente
- $\mathbf{Y_{it}}$ es un vector $1 \times g_2$ de observaciones de g_2 variables endógenas incluidas como covariables, y está permitido que sean correlacionadas con ν_{it};
- $\mathbf{X_{1it}}$ es un vector $1 \times k_1$ de observaciones de variables exógenas incluidas como covariables.
- $\mathbf{Z_{it}} = \mathbf{Y_{it}X_{it}}$;
- $\boldsymbol{\gamma}$ es un vector $g_2 \times 1$ de coeficientes;
- $\boldsymbol{\beta}$ es un vector $k_1 \times 1$ de coeficientes; y

[45] Los datos se pueden encontrar en: http://www.bls.gov/nls/nlsorig.htm
[46] Producto Nacional Bruto.

- $\boldsymbol{\delta}$ es un vector $K \times 1$ de coeficientes, donde $K = g_2 + k_1$.

Para el modelo de efectos fijos del ejemplo se tendría:

$$\ln_w = 1.678 + 0.240\text{tenure} + 0.0118\text{age} - 0.0012\text{age}^2 - 0.0167\text{not}_s\text{msa}$$

donde tenure está en función de union y south. La fracción de la varianza debida al efecto de panel $\sigma_v^2 = 0.7066$ es $\rho = 0.5569$, lo que indica que se debe tomar en cuenta el efecto de panel para estos datos.

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. Las figuras 62 y 63 a continuación se muestran los resultados:

```
. xtivreg ln_w age c.age#c.age not_smsa (tenure = union south), fe
Fixed-effects (within) IV regression          Number of obs      =        19007
Group variable: idcode                        Number of groups   =         4134

R-sq:  within  =      .                       Obs per group: min =            1
       between = 0.1304                                     avg =          4.6
       overall = 0.0897                                     max =           12

                                              Wald chi2(4)       =    147926.58
corr(u_i, Xb)  = -0.6843                      Prob > chi2        =       0.0000

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tenure |   .2403531   .0373419     6.44   0.000     .1671643    .3135419
         age |   .0118437   .0090032     1.32   0.188    -.0058023    .0294897
             |
 c.age#c.age |  -.0012145   .0001968    -6.17   0.000    -.0016003   -.0008286
             |
    not_smsa |  -.0167178   .0339236    -0.49   0.622    -.0832069    .0497713
       _cons |   1.678287   .1626657    10.32   0.000     1.359468    1.997106
-------------+----------------------------------------------------------------
     sigma_u |  .70661941
     sigma_e |  .63029359
         rho |  .55690561   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F  test that all u_i=0:     F(4133,14869) =     1.44      Prob > F    = 0.0000
------------------------------------------------------------------------------
Instrumented:   tenure
Instruments:    age c.age#c.age not_smsa union south
------------------------------------------------------------------------------
```

Figura 62: Regresión de variables instrumentales para datos de panel ajustado en Stata

Note que en este ejemplo se ha modelado a la antigüedad (`tenure`) como una función de las variables `south` y `union`. Es decir que se asume que las variables `south` y `union` *explican* la antigüedad pero no están correlacionadas con el término de error del modelo principal. `south` y `union` son las *variables instrumentales*.

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/nlswork.dta")

library(plm)

## Residuals :
##     Min. 1st Qu.  Median 3rd Qu.     Max.
##  -3.070  -0.325   0.000   0.313   3.640
##
## Coefficients :
##              Estimate  Std. Error t-value  Pr(>|t|)
## age         0.01184373  0.00900322  1.3155    0.1884
## I(age^2) -0.00121445  0.00019684 -6.1696 7.022e-10 ***
## not_smsa -0.01671777  0.03392363 -0.4928    0.6222
## tenure     0.24035305  0.03734191  6.4365 1.260e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1119.5
## Residual Sum of Squares: 5907
## R-Squared      :  0.044277
##       Adj. R-Squared :  0.034638
## F-statistic: -3012.77 on 4 and 14869 DF, p-value: 1
```

Figura 63: Regresión de variables instrumentales para datos de panel ajustado en R

En R se utilizó el comando **`plm`** el cual se basa en el método de mínimos cuadrados. De lo que se observa que se obtienen los mismos resultados en sus coeficientes como en sus parámetros.

Modelos de regresión lineal para errores estándar (xtpcse[47])

Descripción

El objetivo es ajustar un modelo de regresión series de tiempo *sección transversal*. La particularidad en este modelo es que se permite que el término de perturbación (error) sea heterocedástico y correlacionado entre paneles. Funciona como una alternativa viable al método de los mínimos cuadrados generalizados.

Los datos usados para ilustrar el método corresponden a información de 10 empresas en más de 20 años, desde 1935 a 1954; en particular:

Variable	Descripción	Tipo de variable
mvalue	Valor de mercado de la empresa	Numérica
kstock	Valor de su stock de capital	Numérica
invest	Inversión	Numérica

[47] http://www.stata.com/manuals13/xtxtpcse.pdf

Se desea estimar la inversión de cada empresa en función de las demás variables de la tabla anterior. Dada la naturaleza de los datos, es admisible pensar que exista correlación y heterocedasticidad en el error por lo que el método aplica.

Modelo

Se tiene un modelo como:

$$y_{it} = \mathbf{x_{it}}\boldsymbol{\beta} + \epsilon_{it}$$

donde $i = 1, \dots, m$ es el número de unidades (o paneles); $t = 1, \dots, T_i$;T_i es el número de períodos en el panel i; y ϵ_{it} es una perturbación que puede estar correlacionada en el tiempo t o entre los paneles i.

En las estimaciones usadas en el ejemplo se tiene:

$$\text{invest} = -42.71 + 0.1155\text{mvalue} + 0.2306\text{kstock}$$

Cuyos errores estándar son calculados por una correción de panel.

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. Las figuras 64 y 65 a continuación se muestran los resultados:

```
. xtpcse invest mvalue kstock
Linear regression, correlated panels corrected standard errors (PCSEs)
Group variable:   company                     Number of obs      =       200
Time variable:    year                        Number of groups   =        10
Panels:           correlated (balanced)       Obs per group: min =        20
Autocorrelation:  no autocorrelation                         avg =        20
                                                             max =        20
Estimated covariances      =        55        R-squared          =    0.8124
Estimated autocorrelations =         0        Wald chi2(2)       =    637.41
Estimated coefficients     =         3        Prob > chi2        =    0.0000
------------------------------------------------------------------------------
             |             Panel-corrected
      invest |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      mvalue |   .1155622   .0072124    16.02   0.000      .101426    .1296983
      kstock |   .2306785   .0278862     8.27   0.000      .1760225   .2853345
       _cons |  -42.71437   6.780965    -6.30   0.000     -56.00482  -29.42392
------------------------------------------------------------------------------
```

Figura 64: Regresión lineal para errores estándar ajustado en Stata

Nótese que la variable de panel son las empresas y la variable de tiempo son los años. Son paneles balanceados y no se tiene autocorrelación. Además, los errores estándar son consistentes aún cuando los errores de cada observación no sean independientes. Esto implica que los coeficientes estimados van a tener un error estándar *robusto* ante correlaciones entre empresas como paneles y en el tiempo. Los parámetros son estimados por regresión OLS. Ahora, en R se tiene:

```
library(rio)
datos = import("http://www.stata-press.com/data/r13/grunfeld.dta")
```

```
library(panelAR)
datos$time = as.integer(datos$time)
fit1 <- panelAR(invest~mvalue+kstock, timeVar = "time",data = datos, autoCorr='none',
                panelVar = "company", panelCorrMethod = "pcse")
summary(fit1)
```

```
##
## Panel Regression with no autocorrelation and panel-corrected standard errors
##
## Balanced Panel Design:
##  Total obs.:       200 Avg obs. per panel 20
##  Number of panels: 10  Max obs. per panel 20
##  Number of times:  20  Min obs. per panel 20
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -42.714369   6.780965  -6.299 1.91e-09 ***
## mvalue        0.115562   0.007212  16.023  < 2e-16 ***
## kstock        0.230678   0.027886   8.272 1.94e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-squared: 0.8124
## Wald statistic: 637.4061, Pr(>Chisq(2)): 0
```

Figura 65: Regresión lineal para errores estándar ajustado en R

En R se hizo el ajusto con el comando **`panelAR`** el cual también se basa en el método de mínimos cuadrados ordinarios para su implementación. De lo que se observa que se obtienen los mismos resultados en sus coeficientes como en sus parámetros.

Modelos lineales de efectos fijos y aleatorios con perturbaciones AR(1) (xtregar[48])

Descripción

El objetivo es ajustar un modelo de regresión series de tiempo *sección transversal* cuando el término de perturbación es autoregresivo de primer orden. Se debe identificar la variable de panel para los efectos aleatorios. Además, el proceso AR(1) es definido sobre los errores.
Los datos usados para ilustrar el método corresponden a información de 10 empresas desde 1935 a 1954, en particular:

Variable	Descripción	Tipo de variable
mvalue	Valor de mercado de la empresa	Numérica
kstock	Valor de su stock de capital	Numérica
invest	Inversión	Numérica

Se desea estimar la inversión de cada empresa en función de las demás variables de la tabla anterior. Para ello se hace el supuesto de que se tienen efectos aleatorios por empresa y que los errores siguen un proceso AR(1), esto es, dependen del valor anterior. Se aprecia en los resultados de la estimación que los coeficientes son significativos bajo los supuestos mencionados.

Modelo

Se tiene un modelo de la forma:

$$y_{it_{ij}} - \overline{y}_i = \left(\overline{\mathbf{x}}_{\mathbf{it_{ij}}} - \overline{\mathbf{x}}_{\mathbf{i}}\right)\boldsymbol{\beta} + \epsilon_{it_{ij}} - \overline{\epsilon}_i$$

donde

$$\overline{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{it_{ij}} \qquad \overline{x}_{it_{ij}} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{it_{ij}} \qquad \overline{\epsilon}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \epsilon_{it_{ij}}$$

[48] http://www.stata.com/manuals13/xtxtregar.pdf

de modo que el modelo es un proceso AR(1). Usando las estimaciones del ejemplo se tiene:

$$\text{invest} = -63.220 + 0.0949\text{mvalue} + 0.3501\text{kstock}$$

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. Las figuras 66 y 67 a continuación se muestran los resultados:

```
. xtset
       panel variable:  company (strongly balanced)
        time variable:  year, 1935 to 1954
                delta:  1 year

. xtregar invest mvalue kstock, fe
FE (within) regression with AR(1) disturbances  Number of obs      =       190
Group variable: company                         Number of groups   =        10

R-sq:  within  = 0.5927                         Obs per group: min =        19
       between = 0.7989                                        avg =      19.0
       overall = 0.7904                                        max =        19

                                                F(2,178)           =    129.49
corr(u_i, Xb)  = -0.0454                        Prob > F           =    0.0000

------------------------------------------------------------------------------
      invest |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      mvalue |   .0949999   .0091377    10.40   0.000     .0769677     .113032
      kstock |    .350161   .0293747    11.92   0.000     .2921935    .4081286
       _cons |  -63.22022   5.648271   -11.19   0.000    -74.36641   -52.07402
-------------+----------------------------------------------------------------
      rho_ar |  .67210608
     sigma_u |  91.507609
     sigma_e |  40.992469
     rho_fov |   .8328647   (fraction of variance because of u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(9,178) =    11.53                 Prob > F = 0.0000
```

Figura 66: Modelos lineales de efectos aleatorios con perturbaciones ajustado en Stata

El comando de Stata **`xtregar`** ofrece los coeficientes de las variables independientes con sus errores y p-valor. `rho_ar` es el coeficiente del proceso AR(1). `sigma_u` es el valor de la varianza del efecto aleatorio y `sigma_e` es la varianza del error del modelo. rho es el porcentaje de la varianza del efecto aleatorio sobre la suma de `sigma_u + sigma_e.`Por otro lado, en `R` sería:

```
library(nlme)
fit2 <- lme(invest ~ mvalue + kstock, data = datos,
            subset = year!=1944,
               random = ~1| company,
               correlation = corAR1(0, form = ~ year | company))
summary(fit2)
```

```
## Linear mixed-effects model fit by REML
##  Data: datos
##   Subset: year != 1944
##        AIC      BIC    logLik
##   2003.393 2022.779 -995.6963
##
## Random effects:
##  Formula: ~1 | company
##         (Intercept) Residual
## StdDev:    79.22384 70.99542
##
## Correlation Structure: ARMA(1,0)
##  Formula: ~year | company
##  Parameter estimate(s):
##      Phi1
## 0.8069662
## Fixed effects: invest ~ mvalue + kstock
```

```
##                Value Std.Error  DF   t-value p-value
## (Intercept) -42.01446 30.647668 178 -1.370886  0.1721
## mvalue        0.09336  0.008104 178 11.519489  0.0000
## kstock        0.31819  0.031919 178  9.968748  0.0000
##  Correlation:
##        (Intr) mvalue
## mvalue -0.243
## kstock -0.273 -0.136
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.44841992 -0.31349818  0.05425417  0.21496279  3.30283363
##
## Number of Observations: 190
## Number of Groups: 10
```

Figura 67: Modelos lineales de efectos aleatorios con perturbaciones ajustado en R

En R se utilizó el comando **`lme`** el cual también se basa en mínimos cuadrados[49].

Es preciso aclara que este estimador puede manejar paneles desequilibrados y datos desigualmente espaciados. De lo que se ve en los resultados de R y Stata se ve una pequeña variación que depende del método de optimización de cada uno de los comandos.

Modelo tobit para efectos aleatorios en datos de panel (xttobit[50])

Descripción

El objetivo es ajustar un modelo de regresión tobit para datos de panel con efectos aleatorios. Debe declararse una variable que identifique el panel, la variable dependiente y las variables independientes. Generalmente el efecto aleatorio se le atribuye al panel.

Para el ejemplo se usa los datos de una encuesta longitudinal nacional de 1968 en mujeres de 14 a 26 años[51]. Las variables para ajustar el modelo son:

Variable	Descripción	Tipo de variable
ln_wage	Log natural de (salario/deflactor del PNB[52])	Numérica
union	Pertenece a aluna asociación	Factor (Binaria)
age	Edad	Numérica
grade	Grado completo (años de escolaridad completa, entre 1 y 18)	Numérica
not_smsa	Personas que viven fuera de un área estadística metropolitana estándar.	Factor (Binaria)

[49] Stata también usa mínimos cuadrados para estimar los efectos aleatorios.
[50] http://www.stata.com/manuals13/xtxttobit.pdf
[51] Los datos se pueden encontrar en: http://www.bls.gov/nls/nlsorig.htm
[52] Producto Nacional Bruto.

south	1 si es del sur	Factor (Binaria)
year	Año de la entrevista, entre 1968 y 1988	Factor

Se desea explicar el logaritmo del salario (el porcentaje del salario) que se ve afectado por las demás variables. Cabe aclarar que en modelo también se usan las variables `south` y `year` como una interacción (multiplicadas). La variable de panel se llama `idcode`, la cual representa el código de la encuesta (4148 encuestas). Es decir que se puede evidenciar si existe un efecto debido a la variable de panel usando la lógica de un modelo `tobit`. Estos modelos se pueden ajustar en Stata a través del comando **`xttobit`**[53]. En R se puede ajustar este tipo de Comandos utilizando el comando **`CensReg`** del paquete del mismo nombre.

Modelo

Considere el modelo de regresión lineal con efectos por panel de la forma

$$y_{it} = \mathbf{x_{it}}\boldsymbol{\beta} + \nu_i + \epsilon_{it}$$

para $i = 1, \dots, n$ paneles, donde $t = 1, \dots, n_i$. Los efectos aleatorios, ν_i sin i.i.d., $N(0, \sigma_\nu^2)$, y ϵ_{it} es i.i.d. $N(0, \sigma_\epsilon^2)$ independiente de ν_i. Los datos observados pueden ser censurados por la izquiera, derecha o no censurados.

Usando las estimaciones del ejemplo se tiene:

$$\begin{gathered}\ln_w \text{age} = 0.5101 + 0.1430\text{union} + 0.009\text{age} \\ +0.0784\text{grade} - 0.1339\text{not}_s\text{msa} - 0.3507\text{south} \\ -0.0008\text{year} + 0.0031\text{south} * \text{year}\end{gathered}$$

La fracción de la varianza debida al efecto de panel $\sigma_\nu^2 = 0.3045$ es $\rho = 0.599$ y significativa, lo que indica que se debe tomar en cuenta el efecto de panel para estos datos.

[53] El comando xttobit, utiliza el estimador semi paramétrico para Comandos tobit de efectos fijos desarrollado por Honoré (1992)

Aplicación

Se ajusta el modelo descrito en la parte anterior tanto en `Stata` como en `R`. Las figuras 68 y 69 a continuación se muestran los resultados:

```
. use http://www.stata-press.com/data/r13/nlswork3
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)
. xttobit ln_wage union age grade not_smsa south##c.year, ul(1.9)
> intpoints(25) tobit
  (output omitted)
Random-effects tobit regression                 Number of obs      =     19224
Group variable: idcode                          Number of groups   =      4148
Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       4.6
                                                               max =        12
Integration method: mvaghermite                 Integration points =        25
                                                Wald chi2(7)       =   2924.91
Log likelihood  = -6814.4638                    Prob > chi2        =    0.0000
------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |   .1430525   .0069719    20.52   0.000     .1293878    .1567172
         age |    .009913   .0017517     5.66   0.000     .0064797    .0133463
       grade |   .0784843   .0022767    34.47   0.000      .074022    .0829466
    not_smsa |  -.1339973   .0092061   -14.56   0.000    -.1520409   -.1159536
    1.south |  -.3507181   .0695557    -5.04   0.000    -.4870447   -.2143915
        year |  -.0008283   .0018372    -0.45   0.652    -.0044291    .0027725
             |
south#c.year |
          1  |   .0031938   .0008606     3.71   0.000     .0015071    .0048805
             |
       _cons |   .5101968   .1006681     5.07   0.000      .312891    .7075025
-------------+----------------------------------------------------------------
     /sigma_u |   .3045995   .0048346    63.00   0.000     .2951239     .314075
     /sigma_e |   .2488682   .0018254   136.34   0.000     .2452904    .2524459
-------------+----------------------------------------------------------------
          rho |    .599684   .0084097                      .5831174    .6160733
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01)= 6650.63 Prob>=chibar2 = 0.000
  Observation summary:         0  left-censored observations
                           12334     uncensored observations
                            6890 right-censored observations
```

Figura 68: Modelo Tobit para variables instrumentales en datos de panel ajustado en Stata

Note que la opción `ul(1.9)` indica que el máximo del logaritmo de los salarios, con límite superior igual a 1,9. Los datos se tomaron de la base: nlswork de Stata. Se muestran los coeficientes estimados son sus errores y p-valor asociado. `sigma_u` es el valor de la varianza del efecto aleatorio y `sigma_e` es la varianza del error del modelo. rho es el porcentaje de la varianza del efecto aleatorio sobre la suma de `sigma_u + sigma_e.`

```
library("censReg")
library("plm")
pData = pdata.frame(datos, c('idcode'))
pData$south = as.factor(pData$south)

fit1 <- censReg(ln_wage~union + age + grade + not_smsa + south*year,
                data = pData, right = 1.9)
```

```
## Coefficients:
##               Estimate Std. error  t value  Pr(> t)
## (Intercept)  0.5260997  0.1014165    5.188 2.13e-07 ***
## union        0.1475264  0.0067886   21.731  < 2e-16 ***
## age          0.0093073  0.0017352    5.364 8.15e-08 ***
## grade        0.0756414  0.0020175   37.493  < 2e-16 ***
## not_smsa    -0.1381290  0.0089725  -15.395  < 2e-16 ***
## south1      -0.3453784  0.0704965   -4.899 9.62e-07 ***
## year        -0.0002961  0.0018340   -0.161 0.871722
## south1:year  0.0031776  0.0008730    3.640 0.000273 ***
## logSigmaMu  -1.2542640  0.0136738  -91.728  < 2e-16 ***
## logSigmaNu  -1.3721402  0.0071810 -191.078  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Newton-Raphson maximisation, 4 iterations
## Return code 1: gradient close to zero
## Log-likelihood: -6882.449 on 10 Df
```

Figura 69: Modelo Tobit para variables instrumentales en datos de panel ajustado en R

R realiza el ajuste del Comando por máxima verosimilitud asumiendo una distribución Gaussiana de los términos de error. La optimización utiliza el paquete MaxLik, que implementa los siguientes métodos de optimización: Newton Raphson (método por defecto), BHHH, BFGS, SANN o NM. Los coeficientes y las varianzas de los efectos aleatorios y del error estimados por los dos programas presentan ciertas diferencias, atribuibles a la diferencia en los métodos de optimización. Sin embargo las conclusiones que se obtienen de los resultados equivalentes. Además, R reporta solo la verosimilitud del Comando pero Stata reporta adicionalmente como medida de ajuste el resultado del test de Wald. Este test se puede realizar en R con un comando adicional.

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

De los 35 comandos revisados de Stata, 5 de ellos (14%) no tuvieron una equivalencia directa en R. Los mismo son: **`areg`**, el cual aborda los modelos de regresiones lineales de grandes conjuntos en variables binarias. **`boxcox`**, el mismo que maneja modelos de regresión con sus transformaciones. **`eivreg`**, se utiliza para modelos de regresión de errores en las variables. **`etregress`**, maneja modelos de regresión lineal con efectos de tratamiento endógeno. **`ivtobit`** es en cambio para modelos de regresión tobit con variables endógenas.

Los resultados que proporciona Stata por defecto, para poder verlos en R hay que utilizar más instrucciones.

Una de las ventajas de utilizar R es el hecho de que es un software libre por su facilidad de poder consultar el código fuente, se actualiza con frecuencia a más de que sus investigadores ponen a disposición códigos de métodos muy recientes.

De los 35 comandos revisados de Stata, 5 de ellos (14%) no tuvieron una equivalencia directa en R. Los mismo son: **`areg`**, el cual aborda los modelos de regresiones lineales de grandes conjuntos en variables binarias. **`boxcox`**, el mismo que maneja modelos de regresión con sus transformaciones. **`eivreg`**, se utiliza para modelos de regresión de errores en las variables. **`etregress`**, maneja modelos de regresión lineal con efectos de tratamiento endógeno. **`ivtobit`** es en cambio para modelos de regresión tobit con variables endógenas.

El paquete RegUtils es un resultado importante de este proyecto. Debe mencionarse que más que estar enfocado a lograr eficiencia computacional, este facilita la migración o uso paralelo de las rutinas de la tabla 1 para usuarios de R y Stata.

Es preciso mencionar que las equivalencias alcanzadas corresponden a las salidas por defecto de los comandos base. Tanto en R como en Stata, las funciones ofrecen varias opciones. Por ejemplo, “areg” en Stata, ofrece el uso de “fw”. Esta es una opción de ponderar las observaciones si se dispone de factores de expansión. El alcance de este trabajo no cubre ese nivel de detalle, pues, como se mencionó anteriormente, solo coinciden las salidas por defecto.

Analizar el componente teórico de cada uno de los comandos de la tabla 1 escapa del alcance de este trabajo. El lector debe tener claro que se ofrece equivalencias en las salidas de Stata y R, pero adentrarse en la interpretación y aplicación de aquellas es una tarea personal.

Recomendaciones

Utilizar los comandos implementados en R para el manejo de estos modelos ya que no tendría costo para el usuario, además de esto en R se tiene prontitud en el desarrollo de rutinas, acceso al código para modificarlo en función de las necesidades del investigador y facilidades para el trabajo colaborativo al poder compartir análisis sin preocupación de que el receptor disponga de las licencias.

El uso del paquete RegUtils, producto de este trabajo, sin duda es una herramienta muy útil para usuarios migrantes entre R-Stata. Sin embargo, se recomienda que una vez que se alcance el nivel adecuado de familiaridad con R, se explore paquetes y rutinas propias del lenguaje pero que, sin tener las mismas salidas que Stata, logran los mismos objetivos.

Como todos los proyectos realizados en R, se recomienda que todo usuario interesado se motive en generar retroalimentación para su perfeccionamiento. Así, incluso podría ser parte *del Comprehensive R Archive Network* – CRAN.

Es recomendable que los posibles usuarios de estos resultados, exploren la posibilidad de generar analogías en cuanto a la generación de gráficos asociadas a las funciones analizadas.

REFERENCIAS

Los datos de los comandos tanto de R como de Stata son tomados de los manuales dados por el fabricante a través de su página web. *http://www.stata.com/bookstore/functions-reference-manual/*

Libros:

Gujarati, D, and D Porter. 2010. "Econometría (Quinta Edición)." México: McGraw-Hill Interamericana Editores, SA de CV.

Wooldridge, Jeffrey M. 2006. *Introducción a La Econometría: Un Enfoque Moderno*. Editorial Paraninfo.

Baltagi Badi, 2002, "Econometría (Tercera Edición), Editorial: Springer .

Artículos:

Aigner, Dennis, CA Knox Lovell, and Peter Schmidt. 1977. "Formulation and Estimation of Stochastic Frontier Production Function Models." *Journal of Econometrics* 6 (1). Elsevier: 21–37.

Baltagi, Badi H, and Ping X Wu. 1999. "Unequally Spaced Panel Data Regressions with AR (1) Disturbances." *Econometric Theory* 15 (6). Cambridge Univ Press: 814–23.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2014. "Fitting Linear Mixed-Effects Models Using Lme4." *ArXiv Preprint ArXiv:1406.5823*.

Box, George EP, and David R Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 211–52.

Cameroon, CA, and KP Trivedi Microeconometrics Using Stata. 2010. "Revised Edition. StataCorpLp." Stata Press, USA.

Croissant, Yves, Giovanni Millo, and others. 2008. "Panel Data Econometrics in R: The Plm Package." *Journal of Statistical Software* 27 (2): 1–43.

Draper, Norman Richard, Harry Smith, and Elizabeth Pownell. 1966. *Applied Regression Analysis*. Vol. 3. Wiley New York.

Ghalanos, Alexios. 2013. “Introduction to the Rugarch Package.” Version.

Heckman, James J. 1976. “The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models.” In *Annals of Economic and Social Measurement, Volume 5, Number 4*, 475–92. NBER.

Henningsen, Arne, Jeff D Hamann, and others. 2007. “Systemfit: A Package for Estimating Systems of Simultaneous Equations in R.” *Journal of Statistical Software* 23 (4): 1–40.

Lovell, Michael C. 2006. “A Simple Proof of the FWL (Frisch-Waugh-Lovell) Theorem.” *Available at SSRN 887345*.

Meeusen, Wim, and Julien Van den Broeck. 1977. “Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error.” *International Economic Review*. JSTOR, 435–44.

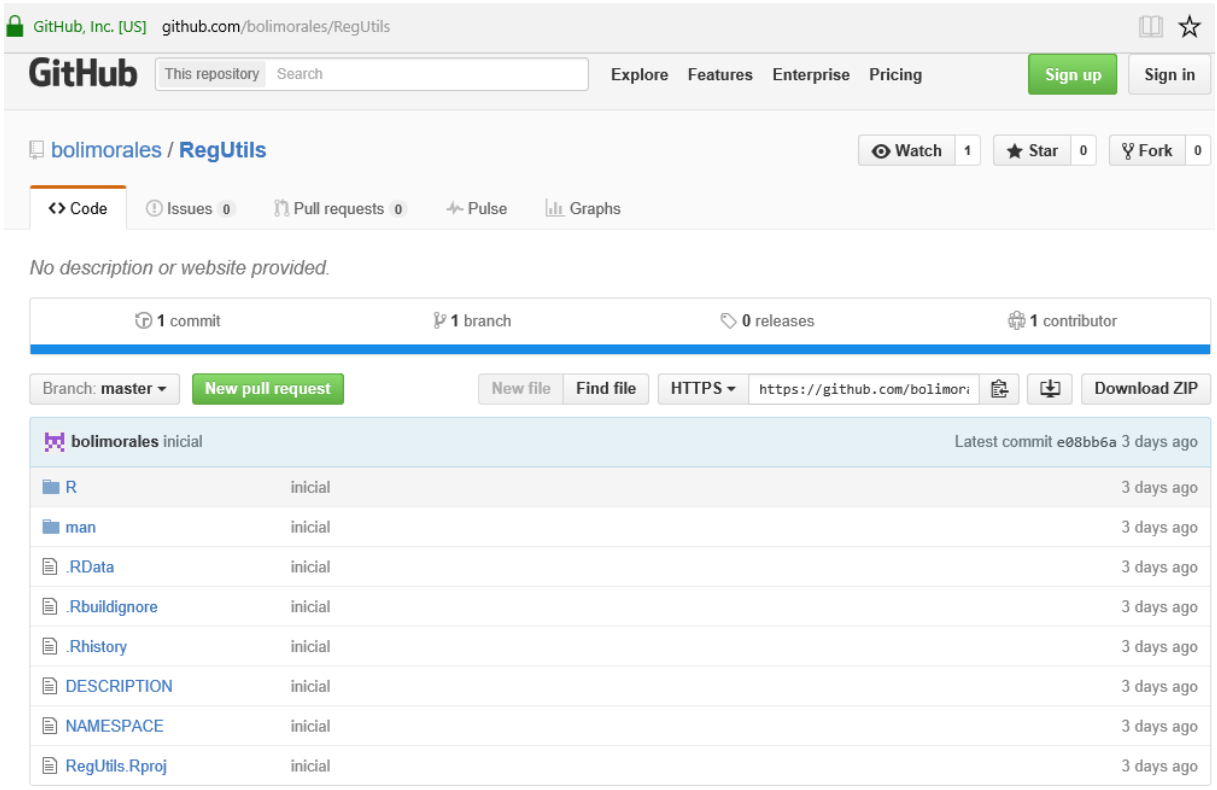
Zellner, Arnold, and Nagesh S Revankar. 1969. “Generalized Production Functions.” *The Review of Economic Studies* 36 (2). Oxford University Press: 241–50.

ANEXOS

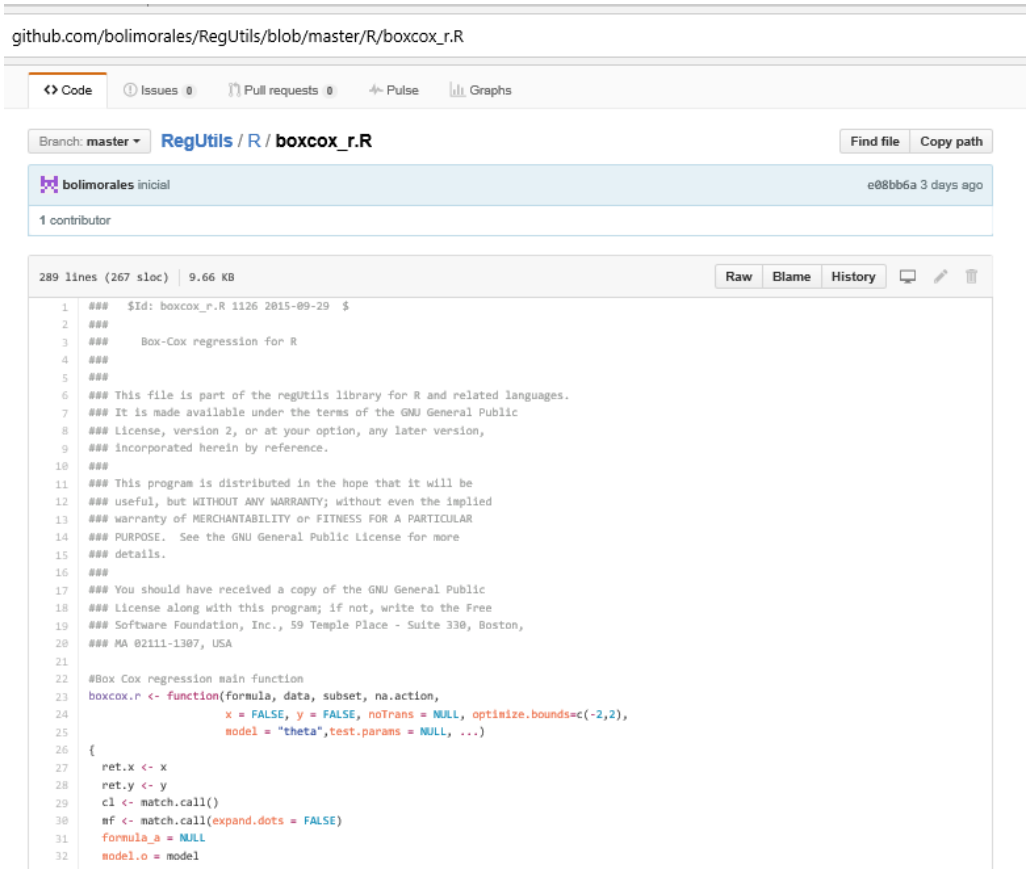
Para poder instalar la librería ***RegUtils*** desde internet se deberían bajar previamente las siguientes librerías: *MaxLik*, *Formula*, *car*, *sandwich* y *censReg* y luego se podría tener acceso con las siguientes líneas de código:

```
library(devtools)
install_github("bolimorales/RegUtils")
library(RegUtils)
```

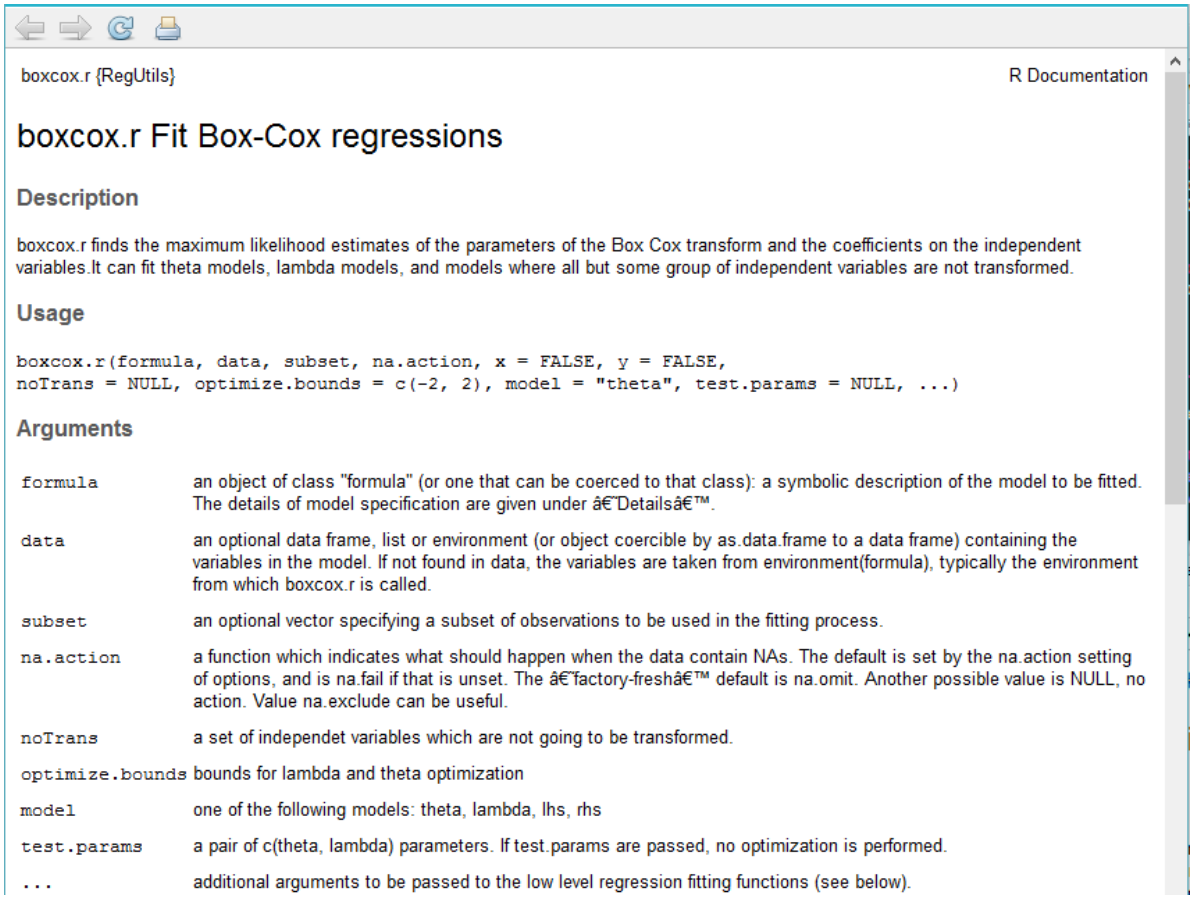
De ese modo ya podría tener acceso desde su computador a los commandos implementados. Otra forma de acceder a la información de RegUtils es a través de internet. Si accede a la página https://github.com/bolimorales/RegUtils, va a encontrarse con una pantalla como la que sigue:



Luego, dentro de la carpeta "R" se encuentra el detalle de las funciones implementadas. Verá la siguiente pantalla:



Recuerde que también puede acceder a la ayuda de la función una vez que haya instalado el paquete. Por ejemplo, luego de cargar la librería, al escribir `help(boxcox.r)` puede ver:



boxcox.r {RegUtils} R Documentation

boxcox.r Fit Box-Cox regressions

Description

boxcox.r finds the maximum likelihood estimates of the parameters of the Box Cox transform and the coefficients on the independent variables.It can fit theta models, lambda models, and models where all but some group of independent variables are not transformed.

Usage

```
boxcox.r(formula, data, subset, na.action, x = FALSE, y = FALSE,
noTrans = NULL, optimize.bounds = c(-2, 2), model = "theta", test.params = NULL, ...)
```

Arguments

formula	an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under â€˜Detailsâ€™.
data	an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which boxcox.r is called.
subset	an optional vector specifying a subset of observations to be used in the fitting process.
na.action	a function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options, and is na.fail if that is unset. The â€˜factory-freshâ€™ default is na.omit. Another possible value is NULL, no action. Value na.exclude can be useful.
noTrans	a set of independet variables which are not going to be transformed.
optimize.bounds	bounds for lambda and theta optimization
model	one of the following models: theta, lambda, lhs, rhs
test.params	a pair of c(theta, lambda) parameters. If test.params are passed, no optimization is performed.
...	additional arguments to be passed to the low level regression fitting functions (see below).

De modo que el usuario puede tener acceso a toda la información que desee.