

# UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN



## PRUEBA DE PHILLIPS-PERRON

**PROFESOR:** *PORRAS CERRON, Jaime Carlos*

**INTEGRANTES:**

- *BULEJE AVILA , Iliana Marihet*
- *CALISAYA MALLCO , Kenia Nora*
- *FLORES CACERES , Fiorella Alexandra*

*La Molina, 2016*

## 1. ASPECTOS GENERALES

Se trata de una prueba de raíz unitaria utilizada en el análisis de series de tiempo, para probar la hipótesis nula de que una serie de tiempo es integrada de orden 1. Se basa en la prueba de Dickey - Fuller ( $\rho = 0$ ) en  $Y_t = \rho Y_{t-1} + \varepsilon_t$ , donde  $\Delta$  es la primera diferencial del operador.

En esta prueba la raíz unitaria fue desarrollada por Phillips y Perrón, que al igual que ADF (Dickey - Fuller aumentado-1988), aborda la cuestión de que el proceso de generación de datos para  $Y_t$  podría tener un orden superior de autocorrelación que es admitido en la ecuación de prueba, invalidando así Dickey-Fuller t-test. Mientras que la prueba de Dickey-Fuller aumentada aborda esta cuestión mediante la introducción de retardos de  $\Delta Y_t$  como variables independientes en la ecuación de la prueba.

Pero la diferencia radica que la prueba ADF, no existe término de diferencia retardada, además PP utilizan métodos estadísticos no paramétricos para evitar la correlación serial en los términos del error, sin añadir términos de diferencia rezagada en la ecuación (esta es la principal diferencia).

## 2. SUPUESTOS

- Las observaciones en una serie de tiempo, son medidos en determinados momentos de tiempo, ordenados cronológicamente (diario, semanal, semestral, anual, entre otros) y, son espaciados entre sí de manera uniforme.
- Las observaciones de la variable de interés deben ser de tipo cuantitativa continua.

## 3. OBJETIVO

Detectar la presencia de una raíz unitaria en una serie de tiempo, probando la hipótesis nula de un proceso con raíz unitaria y una constante, versus la hipótesis alterna de que el proceso generador de los datos es un proceso estacionario alrededor de una tendencia. La raíz unitaria es un indicador de series no estacionarias.

## 4. MARCO TEORICO

### Prueba de Phillips-Perron:

Se basa en la prueba de Dickey-Fuller. Contiene el proceso auto-regresivo AR (1):

$$\Delta Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t$$
$$Y_t = \alpha + \rho Y_{t-1} + \varepsilon_t$$

### Prueba de Hipótesis:

- |                 |   |                                       |
|-----------------|---|---------------------------------------|
| $H_0: \rho = 1$ | o | $H_0: \text{Existe raíz unitaria}$    |
| $H_a: \rho < 1$ | o | $H_a: \text{No existe raíz unitaria}$ |

Mientras que la prueba ADF corrige la correlación serial de orden elevado añadiendo más retardos del término diferenciado de la serie original en el lado derecho de la ecuación, la prueba de Phillips-Perron realiza una corrección del estadístico t sobre el coeficiente  $\rho$  en la regresión AR (1) para considerar la correlación serial en el término  $\varepsilon$ . Mientras que el procedimiento de Dickey-Fuller busca de retener la validez de las pruebas basadas en errores ruido blanco (ampliando el polinomio auto-regresivo al grado  $p+1$ ), el procedimiento de Phillips-Perron modifica las estadísticas para tomar en cuenta distintos tipos de estructura en los errores.

La prueba parte de la estimación por Mínimos Cuadrados Ordinarios y luego el t-estadístico del coeficiente  $\rho$  es corregido.

La distribución asintótica del t-ratio de los parámetros  $\alpha$  en los modelos de la prueba DF depende de la relación  $\frac{\sigma_{\varepsilon}^2}{\sigma^2}$

$$\sigma_{\varepsilon}^2 = \lim_{n \rightarrow \infty} \left[ \frac{\sum_{i=1}^n E(\varepsilon_i^2)}{n} \right] \quad y \quad \sigma^2 = \lim_{n \rightarrow \infty} \left[ \frac{E(\sum_{i=1}^n \varepsilon_i)^2}{n} \right]$$

Se utilizan los residuos obtenidos en la estimación del modelo DF para transformar los estadísticos "t" asociados a los parámetros del mismo. La corrección de las razones "t" "incorrectamente" calculadas, intenta hacerlas independientes de la ratio  $\sigma_{\varepsilon}^2/\sigma^2$ .

La transformación se realizará con estimaciones propuestas para  $\sigma_{\varepsilon}^2$  y  $\sigma^2$  por estos autores, estimaciones para las que se utilizarán los residuos previamente obtenidos en la regresión del modelo DF analizado. Las estimaciones sugeridas son:

$$\hat{\sigma}_{\varepsilon}^2 = \frac{\sum_{i=1}^n e_i^2}{n} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n} + 2 \frac{\sum_{r=1}^l (1 - \frac{r}{l+1}) \sum_{i=r+1}^n e_i e_{i-r}}{n}$$

Se fija un determinado nivel máximo de retardo "l" que se quiere tener en cuenta, por ejemplo "l=5" y se computa para cada uno de esos retardos considerados "r=1,2,3,4..., l" la correlación muestral

$$\sum_{i=r+1}^n e_i e_{i-r}$$

con el número máximo de datos posibles. A continuación, cada una de esas correlaciones muestrales se pondera con el término  $(1-r/l+1)$  dando más importancia a la correlación para un retardo ( $r=1$ ) que a la correlación para retardos más elevados; hecho esto se obtiene la suma ponderada de todas ellas. Las varianzas calculadas en el primer sumando se completan así con el doble de la covarianza muestral calculada siguiendo la expresión matemática básica:

$$E[\varepsilon_i + \varepsilon_j]^2 = E(\varepsilon_i^2) + E(\varepsilon_j^2) + 2E(\varepsilon_i \varepsilon_j)$$

Dado que la expresión para la estimación de  $\sigma^2$  dependerá del valor máximo "l" conviene "testar" la sensibilidad del cálculo a los diferentes valores de éste o bien, seguir la

recomendación de Schwert (1987 y 1989) tomando "l" en función del número de datos según las expresiones:

$$l = \frac{4 \sqrt[4]{n}}{100}; l = \frac{12 \sqrt[4]{n}}{100}$$

Lo cual supone, por ejemplo, l=1 ó l=3 (según la 1ª ó 2ª expresión) para 100 datos, 2 ó 6 para 200 datos, 3 ó 9 para 300 y así sucesivamente (un retardo más por cada 100 datos extra). Una vez computadas las estimaciones de  $\sigma_\varepsilon^2$  y  $\sigma^2$ , corregiremos el valor obtenido para la razón "τ" en la estimación de nuestro modelo según las expresiones:

$$Z(t) = \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}^2} \hat{t} - \frac{\frac{1}{2}(\hat{\sigma}^2 - \hat{\sigma}_\varepsilon^2)}{\hat{\sigma} \sqrt{\frac{\sum_{t=2}^n Y_{t-1}^2}{n^2}}}$$

tanto para el caso de haber estimado el modelo más simple (τ) como en el caso del modelo con constante (τ<sub>μ</sub>), y:

$$Z(t) = \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}^2} \hat{t}_t - \frac{n^3(\hat{\sigma}^2 - \hat{\sigma}_\varepsilon^2)}{4 \hat{\sigma} \sqrt{3 D_y}}$$

para el caso del modelo menos restringido (con término constante y tendencia determinista). En esta última expresión, D<sub>y</sub> se calcula como:

$$D_y = \frac{n^2(n^2 - 1)}{12} \sum_{i=2}^n Y_{i-1}^2 - n \left( \sum_{i=2}^n i Y_{i-1} \right)^2 + n(n+1) \sum_{i=2}^n i Y_{i-1} \sum_{i=2}^n Y_{i-1} - \frac{n(n+1)(2n+1)}{6} \left( \sum_{t=2}^n Y_{t-1} \right)^2$$

Los valores así corregidos de las razones "t" (ó τ) pueden entonces ser comparados sin problemas con las distribuciones tabuladas por Fuller (1976). Por lo tanto, la distribución asintótica del estadístico t de la prueba de Phillips-Perron es la misma que la del estadístico t de la prueba ADF y se contrastan los resultados de la prueba con los valores críticos de MacKinnon (El estadístico t no tiene la distribución convencional de la t-Student. Por ello, Dickey y Fuller simulaban los valores críticos para una selección de distintos tamaños muestrales, simulación ampliada posteriormente por MacKinnon. Estas son las tablas que se utilizan para determinar los valores críticos del test DF.) Si el estadístico t asociado al coeficiente de Y<sub>t-1</sub> es mayor en valor absoluto al valor crítico de MacKinnon, entonces se rechaza la hipótesis nula.

La corrección que realiza esta prueba es no paramétrica debido a que utiliza una estimación del espectro del término ε en la frecuencia cero que es robusta para una forma no conocida de heterocedasticidad y autocorrelación. Utiliza la corrección conocida como de Newey-West para la heterocedasticidad y autocorrelación.

## Aplicación de la Prueba

El caudal es una medida del fluido que circula a través de un ducto. Se obtiene de una relación entre volumen y unidad de tiempo. Las medidas del caudal de un río son importantes para el dimensionamiento de presas, embalses, etc. En función a lo requerido se pueden recoger de forma diaria o anual, y en niveles mínimos, máximos o promedios. Se cuentan con datos acerca del caudal anual máximo de un río (en m<sup>3</sup>/s) desde el año 1917 a 2014. Requerimos probar si esta serie de tiempo es estacionaria a un nivel de significancia de 0.05.

N°	Año	Zt	N°	Año	Zt	N°	Año	Zt	N°	Año	Zt
1	1917	302.3	26	1942	289.8	51	1967	295.1	76	1992	299.7
2	1918	296.0	27	1943	296.4	52	1968	300.3	77	1993	306.5
3	1919	295.4	28	1944	307.1	53	1969	293.3	78	1994	300.9
4	1920	298.9	29	1945	300.5	54	1970	303.2	79	1995	290.4
5	1921	305.8	30	1946	301.7	55	1971	317.6	80	1996	305.1
6	1922	307.3	31	1947	300.5	56	1972	307.2	81	1997	302.6
7	1923	294.6	32	1948	306.0	57	1973	294.5	82	1998	293.8
8	1924	303.8	33	1949	306.1	58	1974	299.7	83	1999	306.4
9	1925	300.3	34	1950	289.3	59	1975	299.8	84	2000	305.3
10	1926	286.3	35	1951	296.8	60	1976	300.7	85	2001	301.0
11	1927	311.9	36	1952	311.4	61	1977	302.9	86	2002	301.0
12	1928	312.1	37	1953	297.7	62	1978	305.6	87	2003	298.3
13	1929	283.7	38	1954	301.3	63	1979	301.5	88	2004	295.8
14	1930	294.6	39	1955	314.1	64	1980	293.9	89	2005	307.8
15	1931	304.8	40	1956	294.2	65	1981	291.9	90	2006	312.4
16	1932	298.2	41	1957	301.5	66	1982	301.0	91	2007	298.9
17	1933	297.9	42	1958	308.5	67	1983	302.3	92	2008	303.2
18	1934	304.9	43	1959	291.5	68	1984	311.1	93	2009	307.7
19	1935	308.8	44	1960	297.6	69	1985	293.5	94	2010	296.2
20	1936	295.9	45	1961	295.9	70	1986	287.4	95	2011	279.1
21	1937	294.2	46	1962	297.5	71	1987	309.1	96	2012	305.6
22	1938	314.3	47	1963	311.9	72	1988	309.1	97	2013	317.6
23	1939	307.5	48	1964	313.9	73	1989	298.3	98	2014	300.2
24	1940	301.8	49	1965	295.9	74	1990	301.4			
25	1941	300.1	50	1966	295.6	75	1991	300.7			

Prueba de Hipótesis:

$H_0: \rho = 1$       o       $H_0$ : Existe raíz unitaria  
 $H_a: \rho < 1$       o       $H_a$ : No existe raíz unitaria

1. Se forman 2 columnas de la siguiente manera:

$Y_t$ : Los valores de  $Z_t$ , exceptuando el primer valor.

$Y_{t-1}$ : Los valores de  $Z_t$ , exceptuando el último valor.

Cantidad de valores de  $Y_t$  :n=97

2. Se crea una columna con valores del 1 al 97, restándole a cada valor 48.5

3. Se forma el modelo lineal:  $Y_t = 1 + tt + Y_{t_1}$  y se hallan los residuales.

```
> res <- lme1(y ~ 1 + tt + Yt1)
> u <- residuals(res)
> u
```

1	2	3	4	5	6	7	8
-4.34296251	-4.98226132	-3.49856464	5.40167246	6.91562658	-5.79220471	3.34267697	-0.13409023
9	10	11	12	13	14	15	16
-14.16209301	11.36754408	11.65693936	-16.75613652	-5.98459331	4.24549780	-2.32723511	-2.66774421
17	18	19	20	21	22	23	24
4.31716276	5.23152011	-4.66662857	-6.43255405	13.66670491	6.91391155	1.17299559	-0.56428265
25	26	27	28	29	30	31	32
-10.86502369	-4.34045970	6.37228353	-0.19843183	0.96108907	-0.24758251	5.23329359	5.34189999
33	34	35	36	37	38	39	40
-11.47187982	-4.06369876	10.56283674	-3.09214624	0.49870113	18.23934186	-6.62290187	0.68293267
41	42	43	44	45	46	47	48
7.59850020	-9.38714215	-3.36860795	-5.05888148	-3.47962232	10.91294462	12.95716076	-5.04865338
49	50	51	52	53	54	55	56
-5.43515348	-5.95024650	-0.76614639	-7.75905078	2.09882640	16.52458320	6.16909435	-6.58674508
57	58	59	60	61	62	63	64
-1.45186340	-1.34476779	-0.45824710	1.73150103	4.42649374	0.32350361	-7.30691975	-3.35146315
65	66	67	68	69	70	71	72
-0.27341449	1.04341485	9.84077673	-7.73760413	-13.52249507	7.73901746	7.81267900	-3.00120374
73	74	75	76	77	78	79	80
0.04134311	-0.66003931	-1.67674006	5.10534291	-0.48110661	-11.01758141	3.62617572	1.17159716
81	82	83	84	85	86	87	88
-7.65237132	4.89824412	3.83519353	-0.46912694	-0.51495711	-3.22823989	-5.75301523	6.22301622
89	90	91	92	93	94	95	96
10.85754512	-2.63777955	1.59387438	6.09733913	-5.39535927	-22.55866643	3.85546434	15.95148049
97							
-1.41338068							

4. Se halla  $\sigma_\varepsilon^2$

$$\sum_{i=0}^{n=97} \frac{u^2}{n} = 52.06429$$

5. Se halla  $\sigma^2 = 27.44789$

6. Se halla el parámetro de retardo  $l = 4\left(\frac{n}{100}\right)^{\frac{1}{4}} = 3.969656$ , pero tomaremos únicamente la parte entera:  $l = 3$

7. Se calcula Dy:

$$n_2 = n^2 = 9409$$

$$trm_1 = n^2(n_2 - 1) \frac{\sum_{t_1}^{97} Y_{t_1}^2}{12} = 6.488116 * 10^{13}$$

$$trm_2 = n * \left( \sum_{n=1}^{97} Y_{t_1} * n \right)^2 = 1.988701 * 10^{14}$$

$$trm_3 = n(n+1) * \left( \sum_{n=1}^{97} Y_{t_1} * n \right) * \left( \sum_{n=1}^{97} Y_{t_1} \right) = 3.97454 * 10^{14}$$

$$trm_4 = \frac{n * (n+1) * (2n+1) * \left( \sum_{n=1}^{97} Y_{t_1} \right)}{6} = 2.634279 * 10^{14}$$

$$Dy = trm_1 - trm_2 + trm_3 - trm_4 = 37265824002$$

8. Se calcula el estadístico Z(t):

	Estimador	Desv.
<b>Intercepto</b>	302.222	31.05273
<b>tt</b>	0.013883	0.032662
$Y_{t_1}$	-0.004034	0.10312

$$Y_{stats} = \frac{(-0.004034 - 1)}{0.10312} = -9.736038$$

$$Stat = \frac{\sqrt{52.06429}}{\sqrt{32.5677}} * \frac{(-9.736038) - (97^3)}{4 * \sqrt{3} * \sqrt{37265824002} * \sqrt{32.5677} * (32.5677 - 52.06429)} = -10.2027$$

9. Se crea la tabla de valores críticos de MacKinnon y se halla el p-valor:

**Tabla:**

	1	2	3	4	5	6	7	8
1	-4.38	-3.95	-3.6	-3.24	-1.14	-0.8	-0.5	-0.15
2	-4.15	-3.8	-3.5	-3.18	-1.19	-0.87	-0.58	-0.24
3	-4.04	-3.73	-3.45	-3.15	-1.22	-0.9	-0.62	-0.28
4	-3.99	-3.69	-3.43	-3.13	-1.23	-0.92	-0.64	-0.31
5	-3.98	-3.68	-3.42	-3.13	-1.24	-0.93	-0.65	-0.32
6	-3.96	-3.68	-3.41	-3.12	-1.25	-0.94	-0.66	-0.33

**TablaT:**

25	50	100	250	500	$1 * 10^5$
----	----	-----	-----	-----	------------

**Tablap:**

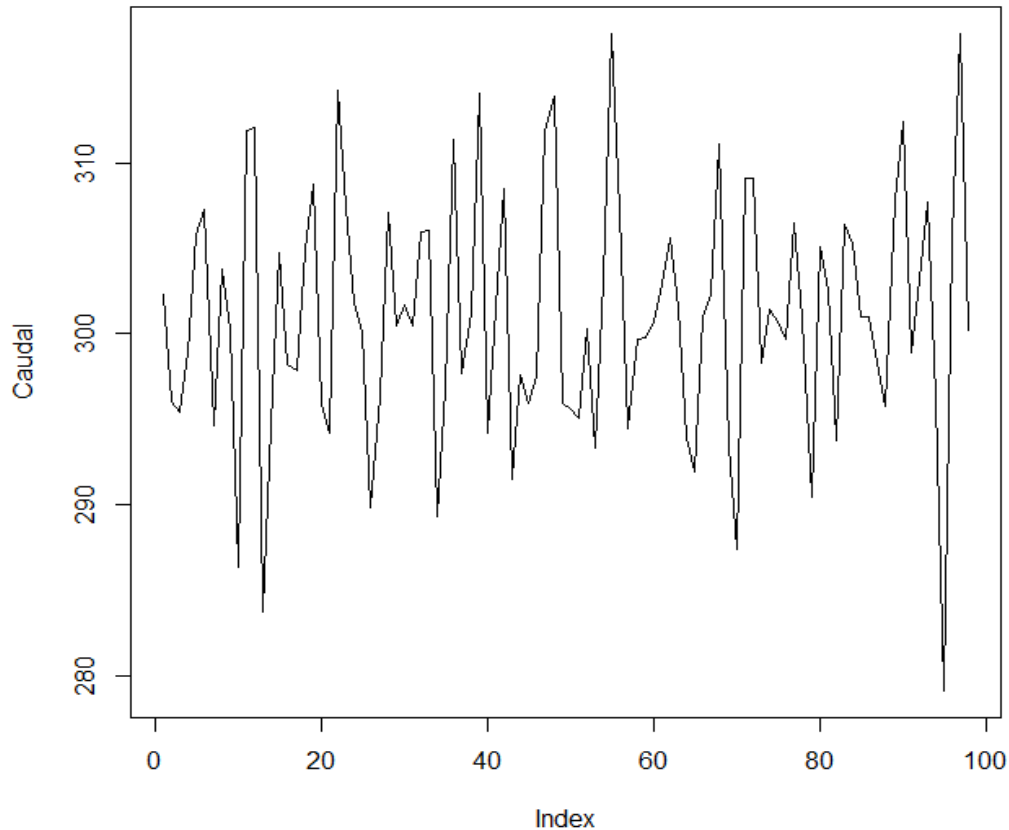
0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99
------	-------	------	-----	-----	------	-------	------

**Tablaip1:**

-4.0466	-3.7342	-3.453	-3.1518	-1.2182	-0.8988	-0.176	-0.2776
---------	---------	--------	---------	---------	---------	--------	---------

$$p_{valor} = \text{aprox}(\text{Tableip1}, \text{Tablep}, \text{Stat}, \text{rule} = 2) \$ y = 0.01$$

El p-valor es 0.01. Por lo tanto, a un nivel de significancia de 0.05 se tiene suficiente evidencia estadística para rechazar la  $H_0$  y afirmar que la serie de tiempo es estacionaria, no presenta raíz unitaria.



### Uso De Un Programa Estadístico Para El Desarrollo De La Aplicación De La Prueba No Paramétrica

A través del programa estadístico R, tenemos para el desarrollo de la aplicación de la prueba:

- La función `PP.test` del paquete `base stats`:

`PP.test(x, lshort = TRUE)`

**Donde:**

<b>X</b>	Vector numérico o serie de tiempo univariada.
<b>lshort</b>	Indica si se utiliza la versión corta o larga del parámetro de retardo de truncamiento.

Se usa la ecuación de regresión general que incorpora una constante y una tendencia lineal y se calcula el estadístico t corregido para un coeficiente de primer orden autorregresivo. Para la estimación de  $\sigma^2$  se utiliza el estimador de Newey-West. Si `lshort` es `TRUE`, el parámetro de retardo de



truncamiento se calcula como  $\text{trunc}(4 * (n / 100) ^ 0.25)$ , caso contrario, como  $\text{trunc}(12 * (n / 100) ^ 0.25)$ . Los p-valores se interpolan de la Tabla 4.2, página 103 de Banerjee et al (1993). Los valores perdidos no se manejan.

- La función **pp.test** del paquete **tseries**

```
pp.test(x, alternative = c("stationary", "explosive"),
type = c("Z(alpha)", "Z(t_alpha)"), lshort = TRUE)
```

**Donde:**

<b>x</b>	Vector numérico o serie de tiempo univariada.
<b>alternative</b>	Indica si la hipótesis alterna es "estacionaria" (por defecto) o "explosiva". Se puede especificar sólo la letra inicial.
<b>type</b>	Indica si se calcula el estadístico "Z (alpha)" (por defecto) o "Z (t_alpha)".
<b>lshort</b>	Indica si se utiliza la versión corta o larga del parámetro de retardo de truncamiento.

Además de calcular el estadístico 'Z(t\_alpha)' para un modelo que incorpora una constante y una tendencia lineal, esta función permite calcular Z(alpha) para un modelo solo con constante. Para la interpolación de los p-valores incluyen a la Tabla 4.1 aparte de la Tabla 4.2 de la página 103 de Banerjee et al (1993). Si el estadístico calculado está fuera de la tabla de valores críticos, se genera un mensaje de advertencia.

```
> caudal<-read.delim("clipboard")
```

```
> head(caudal)
```

```
  Zt
```

```
1 302.3
```

```
2 296.0
```

```
3 295.4
```

```
4 298.9
```

```
5 305.8
```

```
6 307.3
```

```
> caudal<-as.matrix(caudal)
```

```
> PP.test(caudal)
```

```
Phillips-Perron Unit Root Test
```

```
data: caudal
```

```
Dickey-Fuller = -10.203, Truncation lag parameter = 3, p-value = 0.01
```

```
> library(tseries)
```

```
> pp.test(caudal)
```

```
Phillips-Perron Unit Root Test
```

```
data: caudal
```

```
Dickey-Fuller Z(alpha) = -74.465, Truncation lag parameter = 3, p-value  
= 0.01
```

```
alternative hypothesis: stationary Warning message:
```

```
In pp.test(caudal) : p-value smaller than printed p-value  
> pp.test(caudal, type="Z(t_alpha)")
```

Phillips-Perron Unit Root Test

data: caudal

Dickey-Fuller  $Z(t_{\alpha}) = -10.203$ , Truncation lag parameter = 3,  
**p-value = 0.01**

alternative hypothesis: stationary

Warning message:

```
In pp.test(caudal, type = "Z(t_alpha)") :  
p-value smaller than printed p-value
```

El p-valor es 0.01. Por lo tanto, a un nivel de significancia de 0.05 se tiene suficiente evidencia estadística para rechazar la  $H_0$  y afirmar que la serie de tiempo es estacionaria, no presenta raíz unitaria.