## Algunos Tests

### Test de Rachas

Si tenemos observaciones positivas y negativas ordenadas secuencialmente según el tiempo, podríamos preguntarnos si tienen algún patrón particular o si se presentan en forma aleatoria.

Por ejemplo: si tuviéramos la sucesión de residuos siguiente:

$$+ \; + \; - \; + \; - \; - \; - \; - \; + \; + \; - \; + \; + +$$

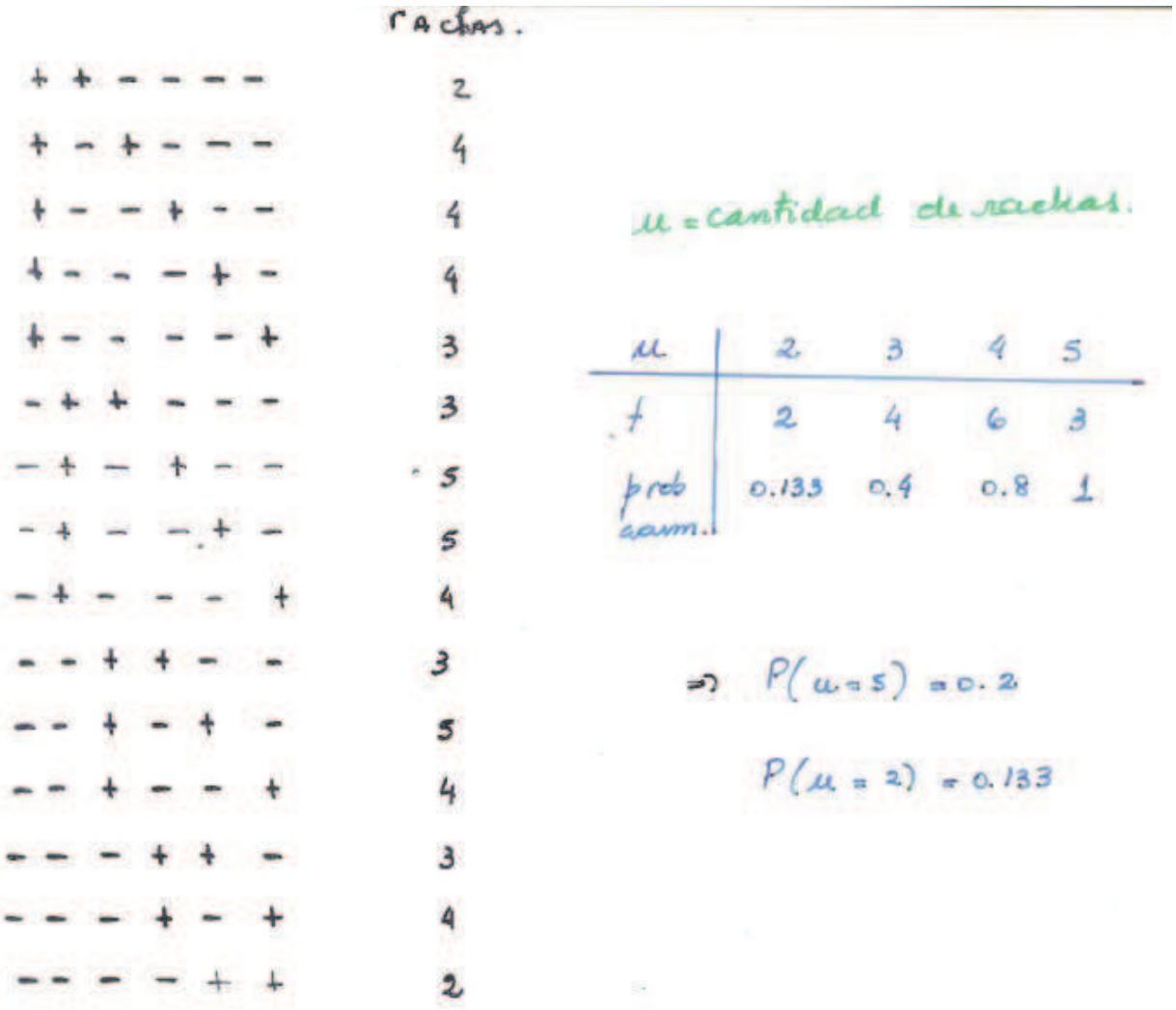
con $n_1 = 8$ residuos positivos, $n_2 = 6$ residuos positivos, $n = 14$ residuos en total y $u = 7$ rachas, ¿hemos observado algo muy poco probable bajo el supuesto de aleatoriedad? ¿Podría haber alguna variable oculta que justifique esto?

Vamos a analizar un caso más sencillo con solo 6 residuos: $2+$ y $4-$.

Un número bajo de rachas hará pensar en una correlación positiva, mientras que un número alto haría sospechar una correlación negativa.

Si $n_1 > 10$ y $n_2 > 10$ puede usarse una aproximación normal para el estadístico del test. Si $n_1 \leq n_2 \leq 10$ se usan las tablas exactas de Sweed y Hasenhart (1943).

El test aproximado resulta de calcular:


rachas.
+ + - - - -
+ - + - - -
+ - - + - -
+ - - - + -
+ - - - - +
- + + - - -
- + - + - -
- + - - + -
- + - - - +
- - + + - -
- - + - + -
- - + - - +
- - - + + -
- - - + - +
- - - - + +
2
4
4
4
3
3
5
5
4
3
5
4
3
4
2
u = cantidad de rachas.
u | 2 | 3 | 4 | 5
f | 2 | 4 | 6 | 3
prob acum. | 0.133 | 0.4 | 0.8 | 1
⇒ P(u=5) = 0.2
P(u = 2) = 0.133

$$Z = \frac{u - \mu \pm 1/2}{\sigma}$$

$$\mu = \frac{2n_1 n_2}{n_1 + n_2} + 1$$

$$\sigma = \frac{2n_1 n_2(2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

(Para muchas chances usar el factor de corrección $-1/2$ y para muy pocas $1/2$)

## Veamos un ejemplo

Consideremos el caso en que examinamos 27 residuos de los cuales 15 son de un signo y 12 son de otro y ordenados secuencialmente de acuerdo con el tiempo presentan 7 rachas. ¿Hay muy pocas rachas?

Supongamos que hubiera $n_1 = 15$ residuos positivos, $n_2 = 12$ residuos positivos, entonces $n = 27$ residuos en total y $u = 7$ rachas, ¿Hay pocas rachas?

$$\mu = \frac{43}{3}$$

$$\sigma = \frac{740}{117}$$

$$Z = \frac{7 - 43/3 + 1/2}{\sqrt{\frac{740}{117}}} = -2{,}713$$

Usando la aproximación normal tenemos:

$$P(Z \leq -2{,}713) \cong 0{,}0033$$

por lo tanto bajo el supuesto de aleatoriedad estaríamos observando un número inusualmente bajo de rachas, por lo tanto rechazaríamos la hipótesis de que las rachas de signos han ocurrido simplemente por azar a los niveles habituales.

## Test de Durbin–Watson

Es un test muy conocido que es útil para detectar cierto tipo de correlación en una serie.

Supongamos que postulamos el modelo:

$$y_i = \beta_o + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{ip-1} + \epsilon_i$$

donde $\epsilon_i \sim N(0, \sigma^2)$ independientes.

En este caso, tenemos que $\rho_s = Corr(y_i, y_{i-s}) = 0 \quad \forall s$.

Supongamos que $\epsilon_i \sim N(0, \sigma^2)$, pero en realidad hay cierta estructura en los errores:

$$\epsilon_i = \rho \epsilon_{i-1} + u_i \qquad \text{Modelo Autorregresivo}$$

donde $\rho$ representa la correlación y $u_i$ las innovaciones, que son independientes de todo el pasado.

Si $\epsilon_i = \rho\epsilon_{i-1} + u_i$ entonces

$$\begin{aligned} Cov(\epsilon_i, \epsilon_{i-1}) &= Cov(\rho\epsilon_{i-1} + u_i, \epsilon_{i-1}) \\ &= \rho\sigma^2 \\ &\Downarrow \\ Corr(\epsilon_i, \epsilon_{i-1}) &= \rho \end{aligned}$$

¿Cuánto vale $Corr(\epsilon_i, \epsilon_{i-s})$? Veamos que $Corr(\epsilon_i, \epsilon_{i-s}) = \rho^s$

Nuestro objetivo es testear:

$$H_0 : \rho_s = 0 \quad v. \quad H_0 : \rho_s = \rho^s$$

para $\rho \neq 0, |\rho| < -1$. Esta alternativa surge del modelo $\epsilon_i = \rho\epsilon_{i-1} + u_i$, donde $u_i \sim N(0, \sigma^2$ e independientes de $\epsilon_{i-1}, \epsilon_{i-1}, \ldots..$ y de $u_{i-1}, u_{i-1}, \ldots...$ Se asume además que la media y la varianza de las $\epsilon_i$ son constantes, más aún: $\epsilon_i \sim N(0, \sigma^2/(1-\rho^2)$

El estadístico del test está basado en los residuos $e_1, \ldots, e_n$:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

¿Cuál es la zona de rechazo? Las tablas de Durbin-Watson proveen para $\alpha = 0{,}05,\ 0{,}0025 y 0{,}01$ valores $d_L$ y $d_U$ para distintos valores de $n$ y de $p$ (cantidad de covariables).
Pueden encontrarse las tablas que usaremos en:

`http://www.imm.bwl.uni-muenchen.de/dateien/3_lehre/market_analysis/durbin_watson_tables.pdf`

**Test de una cola contra alternativas** $\rho > 0$ **de nivel** $\alpha$:

- si $d < d_L \Rightarrow d$ es significativo
- si $d > d_U \Rightarrow d$ no es significativo
- si $d_L \leq d \leq d_U \Rightarrow d$ no hay conclusión

**Test de una cola contra alternativas** $\rho < 0$ **de nivel** $\alpha$:

- idem usando $4 - d$

**Test de una cola contra alternativas** $\rho \neq 0$ **de nivel** $2\alpha$:

- si $d < d_L$ o $4 - d < d_L \Rightarrow d$ es significativo
- si $d > d_U$ y $4 - d > d_U \Rightarrow d$ no es significativo
- en otro caso no hay conclusión

**Veamos un ejemplo extraído de Draper y Smith (1980)**:

Una compañía de gaseosas quiere predecir la venta regional a partir de los gastos mensuales regionales realizados en propagandas. Se dispone de datos de 20 años.

**Data for Soft Drink Concentrate Sales Example**

| | $t$ | (1) Annual Regional Concentrate Sales $y_t$ (units) | (2) Annual Advertising Expenditures $x_t$ ($\times 1000) | (3) Least Squares Residuals $e_t$ | (4) $e_t^2$ | (5) $(e_t - e_{t-1})^2$ |
|---|---|---|---|---|---|---|
| 1960 | 1 | 3083 | 75 | −32.330 | 1045.2289 | |
| 1961 | 2 | 3149 | 78 | −26.603 | 707.7196 | 32.7985 |
| 1962 | 3 | 3218 | 80 | 2.215 | 4.9062 | 830.4771 |
| 1963 | 4 | 3239 | 82 | −16.967 | 287.8791 | 367.9491 |
| 1964 | 5 | 3295 | 84 | −1.148 | 1.3179 | 250.2408 |
| 1965 | 6 | 3374 | 88 | −2.512 | 6.3101 | 1.8605 |
| 1966 | 7 | 3475 | 93 | −1.967 | 3.8691 | 0.2970 |
| 1967 | 8 | 3569 | 97 | 11.669 | 136.1656 | 185.9405 |
| 1968 | 9 | 3597 | 99 | −0.513 | 0.2632 | 148.4011 |
| 1969 | 10 | 3725 | 104 | 27.032 | 730.7290 | 758.7270 |
| 1970 | 11 | 3794 | 109 | −4.422 | 19.5541 | 989.3541 |
| 1971 | 12 | 3959 | 115 | 40.032 | 1602.5610 | 1976.1581 |
| 1972 | 13 | 4043 | 120 | 23.577 | 555.8749 | 270.7670 |
| 1973 | 14 | 4194 | 127 | 33.940 | 1151.9236 | 107.3918 |
| 1974 | 15 | 4318 | 135 | −2.787 | 7.7674 | 1348.8725 |
| 1975 | 16 | 4493 | 144 | −8.606 | 74.0632 | 33.8608 |
| 1976 | 17 | 4683 | 153 | 0.575 | 0.3306 | 84.2908 |
| 1977 | 18 | 4850 | 161 | 6.848 | 46.8951 | 39.3505 |
| 1978 | 19 | 5005 | 170 | −18.971 | 359.8988 | 666.6208 |
| 1979 | 20 | 5236 | 182 | −29.063 | 844.6580 | 101.8485 |
| | | | | $\sum_{t=1}^{20} e_t^2 = 7587.9154$ | | $\sum_{t=2}^{20} (e_t - e_{t-1})^2 = 8195.2065$ |

**Summary Statistics for the Least Squares Model**

| Parameter | Estimate | Standard Error | $t$-Statistic |
|---|---|---|---|
| $\beta_0$ | 1608.508 | 17.0223 | 94.49 |
| $\beta_1$ | 20.091 | .1428 | 140.71 |
| $n=20$ | $R^2=.9991$ | | $MS_E=421.5485$ |

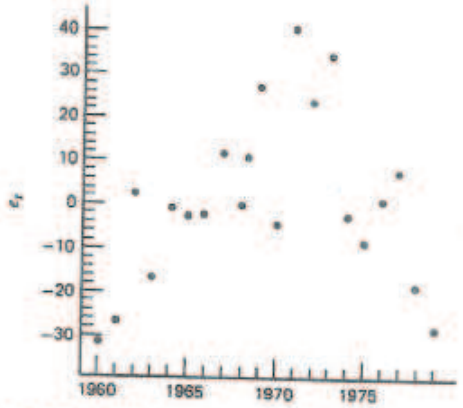



Figure 9.1 Residuals $e_t$ versus time, Example 9.1.

$$H_0: \rho = 0$$

$$H_1: \rho > 0$$

$$d = \frac{\sum_{t=2}^{20} (e_t - e_{t-1})^2}{\sum_{t=1}^{20} e_t^2} = \frac{8195.2065}{7587.9154} = 1.08$$

**Table A.6 Critical Values of the Durbin–Watson Statistic**

| Sample Size | Probability in Lower Tail (Significance Level = $\alpha$) | $k$ = Number of Regressors (Excluding the Intercept) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | | 2 | | 3 | | 4 | | 5 | |
| | | $d_L$ | $d_U$ | $d_L$ | $d_U$ | $d_L$ | $d_U$ | $d_L$ | $d_U$ | $d_L$ | $d_U$ |
| 15 | .01 | .81 | 1.07 | .70 | 1.25 | .59 | 1.46 | .49 | 1.70 | .39 | 1.96 |
| | .025 | .95 | 1.23 | .83 | 1.40 | .71 | 1.61 | .59 | 1.84 | .48 | 2.09 |
| | .05 | 1.08 | 1.36 | .95 | 1.54 | .82 | 1.75 | .69 | 1.97 | .56 | 2.21 |
| 20 | .01 | .95 | 1.15 | .86 | 1.27 | .77 | 1.41 | .63 | 1.57 | .60 | 1.74 |
| | .025 | 1.08 | 1.28 | .99 | 1.41 | .89 | 1.55 | .79 | 1.70 | .70 | 1.87 |
| | .05 | 1.20 | 1.41 | 1.10 | 1.54 | 1.00 | 1.68 | .90 | 1.83 | .79 | 1.99 |
| 25 | .01 | 1.05 | 1.21 | .98 | 1.30 | .90 | 1.41 | .83 | 1.52 | .75 | 1.65 |
| | .025 | 1.13 | 1.34 | 1.10 | 1.43 | 1.02 | 1.54 | .94 | 1.65 | .86 | 1.77 |
| | .05 | 1.20 | 1.45 | 1.21 | 1.55 | 1.12 | 1.66 | 1.04 | 1.77 | .95 | 1.89 |
| 30 | .01 | 1.13 | 1.26 | 1.07 | 1.34 | 1.01 | 1.42 | .94 | 1.51 | .88 | 1.61 |
| | .025 | 1.25 | 1.38 | 1.18 | 1.46 | 1.12 | 1.54 | 1.05 | 1.63 | .98 | 1.73 |
| | .05 | 1.35 | 1.49 | 1.28 | 1.57 | 1.21 | 1.65 | 1.14 | 1.74 | 1.07 | 1.83 |
| 40 | .01 | 1.25 | 1.34 | 1.20 | 1.40 | 1.15 | 1.46 | 1.10 | 1.52 | 1.05 | 1.58 |
| | .025 | 1.35 | 1.45 | 1.30 | 1.51 | 1.25 | 1.57 | 1.20 | 1.63 | 1.15 | 1.69 |
| | .05 | 1.44 | 1.54 | 1.39 | 1.60 | 1.34 | 1.66 | 1.29 | 1.72 | 1.23 | 1.79 |
| 50 | .01 | 1.32 | 1.40 | 1.28 | 1.45 | 1.24 | 1.49 | 1.20 | 1.54 | 1.16 | 1.59 |
| | .025 | 1.42 | 1.50 | 1.38 | 1.54 | 1.34 | 1.59 | 1.30 | 1.64 | 1.26 | 1.69 |
| | .05 | 1.50 | 1.59 | 1.46 | 1.63 | 1.42 | 1.67 | 1.38 | 1.72 | 1.34 | 1.77 |
| 60 | .01 | 1.38 | 1.45 | 1.35 | 1.48 | 1.32 | 1.52 | 1.28 | 1.56 | 1.25 | 1.60 |
| | .025 | 1.47 | 1.54 | 1.44 | 1.57 | 1.40 | 1.61 | 1.37 | 1.65 | 1.33 | 1.69 |
| | .05 | 1.55 | 1.62 | 1.51 | 1.65 | 1.48 | 1.69 | 1.44 | 1.73 | 1.41 | 1.77 |
| 80 | .01 | 1.47 | 1.52 | 1.44 | 1.54 | 1.42 | 1.57 | 1.39 | 1.60 | 1.36 | 1.62 |
| | .025 | 1.54 | 1.59 | 1.52 | 1.62 | 1.49 | 1.65 | 1.47 | 1.67 | 1.44 | 1.70 |
| | .05 | 1.61 | 1.66 | 1.59 | 1.69 | 1.56 | 1.72 | 1.53 | 1.74 | 1.51 | 1.77 |
| 100 | .01 | 1.52 | 1.56 | 1.50 | 1.58 | 1.48 | 1.60 | 1.45 | 1.63 | 1.44 | 1.65 |
| | .025 | 1.59 | 1.63 | 1.57 | 1.65 | 1.55 | 1.67 | 1.53 | 1.70 | 1.51 | 1.72 |
| | .05 | 1.65 | 1.69 | 1.63 | 1.72 | 1.61 | 1.74 | 1.59 | 1.76 | 1.57 | 1.78 |

Source: Adapted from "Testing for Serial Correlation in Least Squares Regression II," by J. Durbin and G. S. Watson, *Biometrika*, Vol. 38, 1951, with permission of the publisher.

## Test de Normalidad de Shapiro–Wilk

Dada una distribución $G_o$, sea $\mathcal{F}$ la familia de diferencias que se obtiene por cambios de posición o escala a partir de $G - o$. Asumiremos que $G - o$ está estandarizada.

Sea $X_1, X_2, \ldots, X_n$ una m.a. con distribución en $\mathcal{F}$, tal que $E(x_i) = \mu$ y $V(x_i) = \sigma^2$.

Consideremos los estadísticos de orden de la muestra:

$$\mathbf{X}_o = (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$$

Por otro lado, sea

$$\mathbf{Z}_o = (Z_{(1)}, Z_{(2)}, \ldots, Z_{(n)})$$

una muestra ordenada de $G_o$, $\mathbf{m} = (m_1, \ldots, m_n)'$ y $V = v_{ij}$, el vector de medias y la matriz de covarianzas de $\mathbf{Z}_o$:

$$m_i = E(Z_{(i)}) \quad v_{ij} = Cov(Z_{(i)}, Z_{(j)})$$

Por lo tanto, para $i = 1, \ldots, n$: $\quad X_{(i)} \simeq \mu + \sigma Z_{(i)}$

En consecuencia, el plot de $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ vs. $(m_1, m_2, \ldots, m_n)$ debería ser aproximadamente lineal.

Una manera de chequear esto es mediante el coeficiente de correlación lineal en este gráfico.

El estadístico del test de Shapiro-Wilk $W$ corresponde a la correlación entre $\mathbf{V}^{-1}\mathbf{m}$ y $\mathbf{X}_o$ para el caso de la familia Normal.

La zona de rechazo es: $W < k_\alpha$

En R la instrucción shapiro.test ejecuta este test devolviendo el p-valor y el estadístico $W$.

```
biomasa<- read.table("C:\\Users\\Ana\\ModeloLineal\\doctex\\biomasa.txt",header=T)
attach(biomasa)
salida<- lm(formula = BIO ~ K + PH)
salida$res
```

```
        1         2         3         4         5         6         7         8         9
-174.95788 -301.36355  390.63567   71.08915 -517.69012 -517.70400  -44.40357  -35.84008 -204.90345
       10        11        12        13        14        15        16        17        18
-271.47716   71.29876  726.37064  618.06946  831.79843  267.83356 -121.24039 -271.03566 -312.78027
       19        20        21        22        23        24        25        26        27
-239.67658 -333.85551 -179.22424 -325.37695 -290.55431 -253.49593 -206.01746  273.70705  -31.03141
       28        29        30        31        32        33        34        35        36
-223.97267 -679.25157  -27.23251 -211.33982  243.45516  782.95205 1135.79900  565.85631 -473.63371
       37        38        39        40        41        42        43        44        45
-241.24364  -55.82630  -95.44412 -102.26077  306.69000  -84.42299   17.49883  264.75622  259.44632
```

```
shapiro.test(salida$res)

        Shapiro-Wilk normality test

data:  salida$res
W = 0.9217, p-value = 0.004813
```