O wykresach kwantylowo-kwantylowych

24 maja 2022

Problem testowania

H_0 : Y_1 ma rozkład $N(\mu, \sigma^2)$, dla pewnego $\mu \in R$ i $\sigma^2 \in (0, \infty)$,
H_1 : Y_1 nie ma rozkładu normalnego

Wykresy normalne

Niech $a = (a_1, ..., a_n)^T$ będzie wektorem pierwszych współrzędnych punktów wykresu, takim że $a^T 1 = 0$, gdzie $1 = (1, ..., 1)^T$. Naszym celem jest dopasowanie prostej do posortowanych rosnąco danych $y = (y_1, ..., y_n)^T$, gdzie $y_1 \leqslant ... \leqslant y_n$, metodą najmniejszych kwadratów.

$$\hat{y}_i = \hat{\mu} + \hat{\sigma} a_i$$

$$\begin{aligned} \hat{\beta} &= \underset{b \in R}{\arg\min} \quad \| Y - Xb \|^2 \\ &= \underset{b \in R}{\arg\min} \quad Y^T Y - 2(X^T Y)^T b + b^T X^T X b \end{aligned}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Wykresy normalne

$$\begin{aligned} \hat{\mu} &= \bar{y}, \\ \hat{\sigma} &= \frac{a^T y}{a^T a} \end{aligned}$$

Do problemu dopasowania prostej regresji na wykresie prawdopodobieństwa można podejść trochę inaczej.
Jeżeli przyjmieny, że $\hat{\sigma} = b^T y$, to po przekształceniu otrzymamy:

$$\begin{aligned} b^T &= \frac{a^T}{a^T a} \\ b^T b &= \frac{a^T a}{(a^T a)^2} = \frac{1}{a^T a} \\ (b^T b)^{-1} b &= a^T a \frac{a}{a^T a} = a \end{aligned}$$

Symetria rozkładu normalnego i jej następstwa

1. $(Z_{(1)}, \ldots, Z_{(n)}) = -(Z_{(n)}, \ldots, Z_{(1)})$, gdzie $Z_{(i)}$ to i-ta staystyka pozycyjna pochodząca ze standardowego rozkładu normalnego.
2. $cov(Z_{(i)}, Z_{(j)}) = cov(-Z_{(n-i+1)}, -Z_{(n-j+1)})$.

Stąd

$$\begin{aligned} m_1 &= -m_n, \\ m_2 &= -m_{n-1}, \\ \vdots & \quad \vdots \end{aligned}$$

Formalny test

Statystyka testowa będzie oparta na rezydualnej sumie kwadratów.

$$\begin{aligned} RSS &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y} - \hat{\sigma} a_i)^2 = \| y - 1\bar{y} - \hat{\sigma} a \|^2 \\ &= (n-1)s^2\{1 - a^T a \frac{\hat{\sigma}^2}{(n-1)s^2}\} \\ &= (n-1)s^2\{1 - r^2(y,a)\}, \end{aligned}$$

gdzie współczynnik korelacji r liczonej na wektorach y i a ma postać:

$$r(y,a) = \left(a^T a \frac{\hat{\sigma}^2}{(n-1)s^2} \right)^{\frac{1}{2}} = \left(\frac{(a^T y)^2}{(n-1)s^2 a^T a} \right)^{\frac{1}{2}},$$

przy czym s^2 jest wariancją próbkową.

Test Shapiro-Wilka

Shapiro i Wilk zaproponowali, aby w celu wyznaczenia estymatorów μ oraz σ skorzystać z uogólnionej metody najmniejszych kwadratów. Zgodnie z teorią wiemy, że najlepszymi liniowymi nieobciążonymi estymatorami tych parametrów są wartości, które minimalizują formę kwadratową

$$(y - \mu 1 - \sigma m)^T V^{-1} (y - \mu 1 - \sigma m).$$

$$\begin{aligned} \hat{\mu} &= \frac{1^T V^{-1} y m^T V^{-1} m}{1^T V^{-1} 1 m^T V^{-1} m} = \frac{1^T V^{-1} y}{1^T V^{-1} 1}, \\ \hat{\sigma} &= \frac{m^T V^{-1} y}{m^T V^{-1} m}. \end{aligned}$$

Test Shapiro-Wilka

Biorąc pod uwagę estymator σ z równania wcześniej, możemy wyznaczyć plotting positions odpowiadające testowi Shapiro-Wilka.

$$b^T = \frac{m^T V^{-1}}{m^T V^{-1} m}.$$

Pozycje teoretycznych statystyk pozycyjnych wynoszą wtedy

$$a^T = (b^T b)^{-1} b^T = \frac{(m^T V^{-1} m) m^T V^{-1}}{m^T V^{-1} V^{-1} m}.$$

Test Shapiro-Wilka

Statystyka testowa W jest zatem zadana wzorem

$$\begin{aligned} W &= \frac{(m^T V^{-1} m) m^T V^{-1}}{m^T V^{-1} V^{-1} m} \cdot \frac{(m^T V^{-1} m) V^{-1} m}{m^T V^{-1} V^{-1} m} \cdot \frac{\hat{\sigma}^2}{(n-1)s^2} \\ &= \frac{(m^T V^{-1} m)^2}{m^T V^{-1} V^{-1} m} \cdot \frac{\hat{\sigma}^2}{(n-1)s^2} \\ &= \frac{(m^T V^{-1} y)^2}{m^T V^{-1} V^{-1} m (n-1)s^2} \\ &= \frac{(c^T y)^2}{(n-1)s^2}, \end{aligned}$$

gdzie

$$c^T = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}}$$

Test Shapiro-Wilka

Aby wyznaczyć a musimy znać wektor wartości oczekiwanych m oraz macierz kowariancji V. Znamy jednak wartości macierzy V tylko dla $n \in \{2, \ldots, 20\}$. Dla $20 < n \leqslant 50$ zaproponowano przybliżenia wektora $d = m^T V^{-1}$ postaci

$$
\begin{aligned}
\hat{d}_i &= 2m_i, \qquad \text{dla } i = 2, 3, \ldots, n-1, \\
\hat{d}_1^{2*} &= \hat{d}_n^{2*} = \begin{cases} \dfrac{\Gamma(\frac{1}{2}n)}{\sqrt{2}\Gamma\{\frac{1}{2}(n+1)\}}, & \text{dla } n \leqslant 20, \\ \dfrac{\Gamma\{\frac{1}{2}(n+1)\}}{\sqrt{2}\Gamma(\frac{1}{2}n+1)}, & \text{dla } n > 20, \end{cases} \\
\hat{d}_1^2 &= \frac{\hat{d}_1^{2*}}{1-2\hat{d}_1^{2*}} \sum_{i=2}^{n-1} \hat{d}_i^{2*}.
\end{aligned}
$$

Test Shapiro-Franci

W 1972 roku Shapiro i Francia wprowadzili test oparty na analizie wariancji dla problemu testowania normalności. Ich statystyka testowa przyjmuje formę

$$W' = \frac{(m^T y)^2}{(n-1)s^2 m^T m}$$

Wtedy

$$\hat{\sigma} = \frac{m^T y}{m^T m}$$

Estymator Bloma

W 1958 roku, Blom zaproponował postać kwantyla, która jest bliska EZ. Rozważamy zależności postaci

$$EZ_i = G\left(\frac{i}{n+1}\right) + R_i, \qquad \text{gdzie}$$

$G(u) = \Phi^{-1}(u)$,
R_i jest błędem, który spełnia nierówność $|R_i| < \frac{M}{n}$, a M nie zależy od i oraz od n. Alternatywnie możemy rozważać także

$$EZ_i = G\left(\frac{i-\alpha}{n-2\alpha+1}\right), \qquad \text{gdzie}$$

$\alpha = \alpha(c)$, a c jest pewną wartością z przedziału od 0 do 1 oraz

Estymator Bloma

Funkcja $\alpha(c)$ maleje do wartości $\frac{\pi}{8}$, a następnie rośnie do $\frac{1}{2}$. W konsekwencji dostajemy nierówność postaci

$$\Phi^{-1}\left(\frac{i-0.39}{n+0.22}\right) \leqslant EZ_i \leqslant \Phi^{-1}\left(\frac{1-0.5}{n}\right).$$

Jako kompromis przyjmujemy, że $EZ_i = \Phi^{-1}\left(\frac{i-\frac{3}{8}}{n+\frac{1}{4}}\right)$.

Test Chen-Shapiro

W przeciwieństwie do wcześniej wymienionych testów, Chen i Shapiro zaproponowali test, w którym w celu wyznaczenia estymatora $\hat{\sigma}$ porównujemy spacje pomiędzy kolejnymi statystykami pozycyjnymi z odległościami pomiędzy ich wartościmi oczekiwanymi pod warunkiem normalności.
Wprowadzamy więc zmienne postaci

$$X_i = \frac{y_{i+1} - y_i}{m_{i+1} - m_i}$$

podobnie jak wcześniej $y_i = \mu + \sigma a_i + e_i = \mu + \sigma m_i + \epsilon_i$, czyli

$$X_i = \frac{\sigma(m_{i+1} - m_i) + \epsilon_{i+1} - \epsilon_i}{m_{i+1} - m_i} = \sigma + \frac{\epsilon_{i+1} - \epsilon_i}{m_{i+1} - m_i} = \sigma + \tilde{\epsilon}_i.$$

Test Chen-Shapiro

$$\hat{\sigma} = \frac{1}{n-1}\sum_{i=1}^{n-1} X_i = \frac{1}{n-1}\sum_{i=1}^{n-1} \frac{y_{i+1}-y_i}{m_{i+1}-m_i}$$

Na początku przekształcamy postać $\hat{\sigma}$ tak, aby przyjmowała formę $b^T y$. Mamy

$$\begin{aligned}
\hat{\sigma} &= \frac{1}{n-1}\sum_{i=1}^{n-1} \frac{y_{i+1}-y_i}{m_{i+1}-m_i} \\
&= \frac{1}{n-1}\left[\frac{y_2}{m_2-m_1} - \frac{y_1}{m_2-m_1} + \ldots + \frac{y_{i+1}}{m_{i+1}-m_i} - \frac{y_i}{m_{i+1}-m_i}\right] \\
&= \frac{1}{n-1}\left[\frac{-y_1}{m_2-m_1} + \frac{y_2(m_3-2m_2+m_1)}{(m_2-m_1)(m_3-m_2)} + \ldots + \frac{y_i(m_{i+1}-2m_i+m_{i-1})}{(m_i-m_{i-1})(m_{i+1}-m_i)} + \frac{y_{i+1}}{m_{i+1}-m_i}\right] \\
&= b^T y,
\end{aligned}$$

gdzie

$$b_i = \begin{cases} \frac{1}{n-1}\cdot\frac{-1}{m_2-m_1}, & \text{dla } i = 1, \\ \frac{1}{n-1}\cdot\frac{m_{i+1}-2m_i+m_{i-1}}{(m_i-m_{i-1})(m_{i+1}-m_i)}, & \text{dla } i = 2,3,\ldots,n-1, \\ \frac{1}{n-1}\cdot\frac{1}{m_n-m_{n-1}}, & \text{dla } i = n. \end{cases}$$

The new normal plot

Problem RSS zastąpimy teraz przez zagadnienie minimalizacji odległości Wassersteina pomiędzy empiryczną funkcja kwantylową i najlepiej dopasowaną do niej teoretyczną funkcją kwantylową, to znaczy

$$\arg\min_{\mu,\sigma} = \int_0^1 \{F_n^{-1}(t) - \mu - \sigma\Phi^{-1}(t)\}^2 \, dt.$$

Empiryczna funkcja kwantylowa zadana jest jako

$$\begin{array}{lrcll} & F_n^{-1}(t) & = & y_i & \text{dla} \quad \frac{i-1}{1} < y < \frac{i}{n} \\ \text{oraz} & F_n^{-1}(0) & = & y_1. & \end{array}$$

The new normal plot

$$a^T = (a_1, \ldots, a_n), \quad \text{gdzie} \quad a_i = \frac{\varphi_{i-1} - \varphi_i}{\sum(\varphi_{i-1} - \varphi_i)^2}, \quad \text{dla} \quad i = 1, \ldots, n.$$

Test oparty na statystyce testowej

$$R_n = \frac{s_n^2 - \hat{\sigma}_n^2}{s_n^2} = 1 - \frac{\hat{\sigma}_n^2}{s_n^2},$$

charakteryzuje się podobna mocą co test Shapiro-Wilka.

Tabela podsumowująca

Tabela: Porównanie plotting positions dla rozpatrywanych w pracy testów z wartościami optymalnymi (n=10).

m	a S-W	$\Phi^{-1}(\frac{i-\frac{3}{8}}{n+\frac{1}{4}})$	$\Phi^{-1}(\frac{i-\frac{1}{2}}{n})$	$\Phi^{-1}(\frac{i}{n+1})$	a C-S	New
-1.539	-1.655	-1.547	-1.645	-1.335	-1.613	-1.829
-1.001	-0.984	-1.000	-1.036	-0.908	-0.940	-1.089
-0.656	-0.6394	-0.655	-0.674	-0.605	-0.594	-0.706
-0.376	-0.366	-0.375	-0.385	-0.349	-0.337	-0.403
-0.123	-0.119	-0.123	-0.126	-0.114	-0.109	-0.131
0.123	0.119	0.123	0.126	0.114	0.109	0.131
0.376	0.366	0.375	0.385	0.349	0.337	0.403
0.656	0.639	0.655	0.674	0.605	0.594	0.706
1.001	0.984	1.000	1.036	0.908	0.940	1.089
1.539	1.655	1.547	1.645	1.335	1.613	1.829