MA 575 Linear Models:

Cedric E. Ginestet, Boston University

Testing for Normality using Q-Q Plots

Week 12, Lecture 2

1 Theory of Q-Q Plots

Quantile vs. quantile (Q-Q) plots permit to evaluate the 'closeness' of two cumulative distribution functions (CDFs). The advantage of this approach is that such comparisons can be conducted regardless of whether we are considering empirical or theoretical distributions.

The empirical CDF, which is referred to as the EDF, and is based on n realizations for some random variables X, such that $X_i \sim F$, for $i = 1, \ldots, n$, will be constructed as follows,

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}\{X_i \leq x\}.$$

for any $x \in \mathbb{R}$. By the strong law of large numbers, the EDF converges *pointwise* to the CDF almost surely, such that

$$F_n(x) \to F(x), \qquad \text{a.s.},$$

for every $x \in \mathbb{R}$, and where $F(x) := \mathbb{P}[X \leq x]$. That is, the event $\{\omega \in \Omega : |F_n(x;\omega) - F(x;\omega)| \to 0\}$ has probability one. Formally, provided that we have constructed a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where the variables X_i's and X are defined, we then have

$$\mathbb{P}\left[\left\{\omega \in \Omega : \lim_{n\to\infty} |F_n(x;\omega) - F(x;\omega)| = 0\right\}\right] = 1,$$

for every $x \in \mathbb{R}$.

1.1 Quantile Distribution

When the function $F : \mathbb{R} \mapsto [0, 1]$ is (i) *continuous* and (ii) *strictly monotonic*, we can define

$$Q(p) := F^{-1}(p),$$

for some real number $p \in [0, 1]$. That is, $\forall\ p \in [0, 1]$, there exists a unique $x \in \mathbb{R}$, such that $F(x) = p$, and we define this x as $Q(p) = F^{-1}(p)$.

In general, however, CDFs need not be continuous and strictly monotonic. The sole condition, which is true for all CDFs is càdlàg ("*continue à droite, et limite à gauche*"), which means that for every $x_0 \in \mathbb{R}$,

$$\exists \lim_{x \uparrow x_0} F(x);$$
$$\exists \lim_{x \downarrow x_0} F(x) = F(x_0).$$

Therefore, using this property, it is now possible to define the quantile function of any random variable, in the following manner,

$$Q(p) := \inf \{x \in \mathbb{R} : F(x) \geq p\},$$

for every $p \in [0, 1]$. Like the CDF, the QDF is monotone non-decreasing in its argument. It is continuous for continuous random variables, and discrete for discrete random variables. However, whereas the $F(x)$ is right-continuous, its inverse is, by convention, *continuous from the left* and has limit on the right. That is, the QDF is càglàd ("*continue à gauche, et limite à droite*"), which means that for every $x_0 \in \mathbb{R}$,

$$\exists \lim_{x \uparrow x_0} Q(x) = Q(x_0);$$
$$\exists \lim_{x \downarrow x_0} Q(x).$$

1.2 Empirical Quantiles

The (theoretical) quantile distribution function (QDF) of a particular random variable is readily obtained as the inverse of its EDF. For some set of realizations X_i's, of size n, we define the empirical QDF as follows,

$$Q_n(p) := \min \{X_1, \ldots, X_n : F_n(X_i) \geq p\},$$

where F_n is the EDF of the random vector as defined in equation (1). For notational convenience, the quantile of a particular random variable will be denoted by

$$X_{(p)} := Q_n(p).$$

2 Testing Normality in Linear Models

2.1 Q-Q Plots of Residuals

1. We first sort the residuals $\widehat{e}_i$'s in increasing order, such that we obtain the *sample order statistics*,

   $$z_{(1)} \leq z_{(2)} \leq \ldots \leq z_{(n)}.$$

   This is a *partial order*.

2. Next, consider a standard normal sample of size n, and let $u_{(j)}$ denote the j^{th} *expected order statistics*, such that

   $$u_{(j)} \doteq \Phi^{-1}\left(\frac{j - 3/8}{n + 1/4}\right),$$

   where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal CDF (see Bloom, 1958).

3. If the z_i's are normally distributed as assumed in standard regression, then we should obtain a **linear** relation between the ordered residuals and the expected residuals under the Normal assumption,

   $$\mathbb{E}[z_{(i)}] = \mu + \sigma u_{(i)},$$

   so that the plotting of $z_{(i)}$ and $u_{(i)}$ will be a **straight line**. That is, we are then considering the 45° line.

2.2 Expected Order Statistics

The quantity of interest can be formulated as follows,

$$\mathbb{E}_{\Phi}[X_{(j)}].$$

Consider the difficulty of estimating such an integral. For a given $1 < j < n$, let a realization of $X_{(j)}$ be simply denoted by x, then this integral is based on:

- $j-1$ realizations with probability $F_X(x)$, with indices $i_1 = 1, \ldots, j-1$;
- $n-j$ realizations with probability $1 - F_X(x)$, with indices $i_2 = j+1, \ldots, n$;
- and the main realization x occurring with probability $f_X(x)$.

Putting all the elements together, we obtain the following density function evaluated at the point $x \in \mathbb{R}$,

$$f_j(x) = \binom{n}{1}\binom{n-1}{j-1} f_j(x)[1 - F(x)]^{n-j} F(x)^{j-1}.$$

Now, it suffices to integrate this particular quantity in order to obtain the desired expected order statistic, such that

$$\mathbb{E}_{\Phi}[X_{(j)}] := \int_{\mathbb{R}} x f_j(x) dx.$$

But this quantity can be approximated using

$$\mathbb{E}_{\Phi}[X_{(j)}] \doteq \mu + \Phi^{-1}\left(\frac{j-\alpha}{n-2\alpha+1}\right) = \Phi^{-1}\left(\frac{j-3/8}{n-1/4}\right),$$

assuming that $\mu = 0$, and where $\alpha := 3/8$.