Chapiter 5 Normality test.

Normality test

$$e = M\varepsilon = (1-H)\varepsilon, \quad H = X(X'X)^{-1}X'$$

$$H = \begin{pmatrix} h_1' \\ h_2' \\ \vdots \\ h_n' \end{pmatrix} \qquad e_i = \varepsilon_i - h_i'\varepsilon$$

$$h_i' h_i = h_{ii}$$

$$var(h_i'\varepsilon) = h_i' h_i \, var(\varepsilon) = \sigma^2 h_{ii}$$

$$h_{ii} = x_i'(X'X)^{-1}x_i$$

$h_i'\varepsilon \xrightarrow{P} 0$ when $n \to \infty$ and $h_{ii} \to 0$

proof: $h_{ii} \to 0$, $\Rightarrow var(h_i'\varepsilon) \to 0$

$$P(|h_i'\varepsilon| \geq \eta) \leq \frac{var(h_i'\varepsilon)}{\eta^2}$$

$$\Rightarrow 0 \leq \lim_{n\to\infty} P(|h_i'\varepsilon| \geq \eta) \leq \lim_{n\to\infty} \frac{var(h_i'\varepsilon)}{\eta^2}$$

when h_{ii} is small, we can use e_i instead of ε_i

Graphic analysis : Normal probability plot.

Let $z_{(1)} < z_{(2)} \cdots z_{(n)}$ be values arranged in ascending order of the random variables. independent and identically distributed $z_1, \ldots z_n$ each obeying a centered reduced normal law : $z_i \sim N(0,1)$

$$E(z_i) \approx \gamma_i = \Phi^{-1}\left[\frac{(i-\frac{3}{8})}{(n+\frac{1}{4})}\right]$$

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{t^2}{2}\right) dt.$$

Let $u_{(1)} < \cdots < u_{(n)}$ be values arranged in ascending order of the random variables. $u_i \sim N(\mu, \sigma^2)$

$$z_{(i)} = \frac{u_{(i)} - \mu}{\sigma^2} \sim N(0,1)$$

$$E(z_{(i)}) = E\left(\frac{u_{(i)} - \mu}{\sigma^2}\right) = \gamma_i$$

$$E[u_{(i)}] \approx \mu + \sigma \gamma_i$$

The graph of U(i) versus γ_i should be approximately a straight line. This type of graph is called a rankit plot or normal probability plot (NPP)

To test the normality of the errors, e(i) is plotted against γ_i where e(i) are values of residuals (or studentized residuals or preferably studentized residuals by cross-validation).

When the graph looks like a straight line, the residuals should obey approximately to a normal distribution, and therefore we might conclude that the ε_i (errors) obey approximately to a normal distribution.

(internal) studentized residuals :

$$e_i^{(s)} = \frac{e_i}{s\sqrt{1-h_{ii}}}$$

(external) studentized residuals (Rstudent) :

$$e_i^{*} = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$$

$s_{(i)}$: estimator of σ in the model of i^{th} observation.

$$var(e_i) = \sigma^2(1-h_{ii})$$

$$\Rightarrow var(e_i^{(s)}) \cong 1.$$

Shapiro-Wilk test

Let $u_1, u_2, \dots u_n$ be independently and identically distributed and assume that $u_{(1)} < \dots < u_{(n)}$ are their ordered values.

Set $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (u_i - \bar{u})^2$ $\quad \bar{u} = \frac{1}{n} \sum_{i=1}^{n} u_i$

the Shapiro-Wilk test statistic is given by

$$W = \sum_{i=1}^{n} \frac{a_i u_{(i)}}{s}$$

where $a_1, \dots a_n$ depend on the expected values of the order statistics from a standard normal distribution and are tabulated.

The null hypothesis of normality is rejected if $W \leq W_\alpha$. where W_α is a tabulated critical point.

The Shapiro-Wilk statistic is used when the sample size is less than 50.

The test statistic W takes values between 0 and 1, with values close to 1 indicating near-normality. The residuals e_i replace the u_i's in usual applications of Shapiro-Wilk test to regression.

* normal plot in SAS procedure PROC UNIVARIATE

  Check W statistic.

Kolmogou - Smirnou Test

used when sample size is larger than 50.

A set of observations $u_1, \dots u_n$ come from any specified distribution function $F_H(x)$.

Let $F(x)$ be the empirical distribution of u_i's,

i.e. $F(x) = \frac{u_x}{n}$ where u_x is the number of u_i's that are not greater than x.

Kolmogorou statistic:

$$D = \sup_x |F(x) - F_H(x)|$$

* SAS: statistic D.

The hypothesis that the u_i's have the distribution given by F_H is rejected for large values of D. ($D \geq D_\alpha$)

In our case, $F_H(x)$ is the distribution function of the normal distribution. The mean is the same as that of the residuals (which is zero if there is a constant term) and the variance is s^2.

Asymptotic Theory.

Theorem Gnedenko and Kolmogorov.

Let $Z_1, Z_2, \ldots Z_n$ be independently and identically distributed with mean zero and variance σ^2.

For $i=1,\ldots n$, let $\{a_{ni}\}$ be a sequence of constants such that

$$\max_{1\le i\le n} |a_{ni}| \to 0 \quad \text{and} \quad \sum_{i=1}^{n} a_{ni}^2 \to 1 \quad \text{as } n\to\infty.$$

Then, as $n\to\infty$, $\sum_{i=1}^{n} a_{ni} Z_i \to N(0,\sigma^2)$, i.e.

$a_{ni}Z_i$ converges to a random variable which has a normal distribution with mean 0 and variance σ^2.

For example, in simple linear regression,

$$\hat{\beta}_1 - \beta_1 = \sum_{i=1}^{n} a_{ni}\,\varepsilon_i \to N \quad \text{when}$$

$$\max_{1\le i\le n} |a_{ni}| \to 0. \quad \text{and} \quad \sum_{i=1}^{n} a_{ni}^2 \to 1. \quad \text{or.}$$

$$\text{when} \quad \max_{1\le i\le n} h_{ii} \to 0.$$

Invoking Large Sample Theory.

Let $\hat{\beta}$ be the least squares estimate of β in the usual multiple regression model and assume the Gauss-Markov Conditions hold.

If in addition, the observations are independent

And $\max_{1\le i\le n} h_{ii} \to 0,$

where h_{ii} are diagonal elements of the matrix

$X(X'X)^{-1}X'$, then

$$\frac{(C\hat{\beta} - C\beta)' [C(X'X)^{-1}C']^{-1} (C\hat{\beta} - C\beta)}{s^2} \to \chi^2_r .$$

as $n \to \infty$. ⊛.

Where C is an $r \times (k+1)$ matrix of rank $r \leq k+1$.

N.B. $\frac{(\hat{\beta} - \beta)' X'X (\hat{\beta} - \beta)}{\sigma^2} \to \chi^2_{k+1}$

However, we suggest using

$$\frac{1}{rs^2} (C\hat{\beta} - C\beta)' [C(X'X)^{-1}C']^{-1} (C\hat{\beta} - C\beta) \sim$$

$F_{r, n-k-1}$ ** in place of ⊛ since it appears to yield a better approximation.

Notice that the statistic given in ** is the one on which all our tests and confidence regions are based.

Thus, all these procedures may be used without change if the h_{ii}'s are small.