– To complement the graphical methods just considered for assessing residual normality, we can perform a hypothesis test in which the null hypothesis is that the errors have a normal distribution. A large p-value, and hence a failure to reject this null hypothesis, is a good result. It indicates that it is reasonable to assume that the errors have a normal distribution.

– Typically, assessment of the appropriate residual plots is sufficient to diagnose deviations from normality. However, a more rigorous and formal quantification of normality may sometimes be requested. For each test discussed below, the formal hypothesis test is written as:

$$H_0 : \text{errors are normally distributed} \quad \text{vs.} \quad H_a : \text{errors are not normally distributed.}$$

– While hypothesis tests are usually constructed to reject the null hypothesis, this is a case where we actually hope to *fail to reject H_0*, since that would mean the errors follow a normal distribution.

The Anderson-Darling Test

– The Anderson-Darling Test measures the area between a fitted line (based on the chosen distribution) and a nonparametric step function (based on the plot points). The statistic is a squared distance that is weighted more heavily in the tails of the distribution. Smaller Anderson-Darling values indicate that the distribution fits the data better. The test statistic is given by

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left[(2i-1)\left(\ln F(Y_i) + \ln(1 - F(Y_{n+1-i})) \right) \right],$$

where F is the cumulative distribution function (CDF) of the normal distribution.

– The test statistic A^2 is compared against critical values from the normal distribution in order to determine the corresponding p-value. If the data are approximately normal, the ECDF closely follows F, and A^2 is small. Large values of A^2 indicate strong departures from normality, leading to rejection of the null hypothesis.

The Shapiro-Wilk Test

– The Shapiro-Wilk Test uses the test statistic

$$W = \frac{\left(\sum_{i=1}^{n} a_i Y_{(i)} \right)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2},$$

where $Y_{(i)}$ denotes the ith order statistic, and the coefficients a_i are calculated using the means, variances, and covariances of the order statistics of a sample from a standard normal distribution.

– The statistic W is compared against tabulated critical values from the distribution of W. The numerator represents the squared correlation between the sample order statistics and their expected normal values, while the denominator represents the total variation in the sample. Thus, W is close to 1 when the sample data follow a straight line in a normal QQ plot. Small values of W occur when the sample deviates from normality (e.g., due to skewness or heavy tails), causing weaker alignment between observed and expected values. Consequently, small W values lead to rejection of the null hypothesis of normality.

The Kolmogorov-Smirnov (Lilliefors) Test

– The Kolmogorov-Smirnov Test (also known as the Lilliefors Test) compares the empirical cumulative distribution function (ECDF) of the sample data with the theoretical cumulative distribution function expected if the data were normally distributed. If this observed difference is sufficiently large, the test will reject the null hypothesis of population normality. The test statistic is given by

$$D = \max_{1 \le i \le n} \left| F(Y_{(i)}) - \frac{i}{n} \right|,$$

where $F(Y_{(i)})$ is the CDF of the normal distribution evaluated at the ith ordered Y.

– The test statistic D is compared against the critical values from the normal distribution in order to determine the corresponding p-value. If the sample data follow the assumed distribution, the two CDFs will be close together, producing a small D value. Large values of D indicate strong deviations between the ECDF and theoretical CDF, leading to rejection of the null hypothesis of normality. The Lilliefors test adjusts the critical values of the KS test to account for the fact that the mean and variance are estimated from the data.

Summary of Normality Tests

Each of the above tests provides a formal mechanism for assessing whether the residuals are approximately normally distributed:

- **Anderson–Darling Test:** Places more weight on the tails of the distribution and is therefore more sensitive to deviations from normality in the tails than the Shapiro–Wilk Test.

- **Shapiro–Wilk Test:** Generally regarded as the most powerful test for detecting departures from normality, especially for small to moderate sample sizes. It focuses on the correlation between the data and the corresponding normal scores.

- **Kolmogorov–Smirnov (Lilliefors) Test:** Compares the empirical and theoretical cumulative distributions and is more general but tends to be less powerful for detecting subtle departures from normality.

– For example, applying the above three tests to test the normality assumption of the model fitted to the Salary1 dataset, we get the following results:

```
ad.test(res)
A = 0.63246, p-value = 0.09379


shapiro.test(res)
W = 0.95523, p-value = 0.05616


lillie.test(res)
D = 0.10512, p-value = 0.1817
```

- The Anderson-Darling test ($A = 0.632$, $p = 0.094$) does not show strong evidence against normality. It places more weight in the tails, and the relatively high p-value suggests the tails are consistent with a normal distribution.

- The Shapiro-Wilk test ($W = 0.955$, $p = 0.056$) gives the smallest p-value among the three, but it still exceeds 0.05, suggesting that any deviation from normality is minor and not statistically significant.

- The Lilliefors test ($D = 0.105$, $p = 0.182$) supports the same conclusion that the residual distribution does not significantly deviate from normality.

Example 1

Apply above three tests to test the normality assumption of the model fitted to the NYC dataset.