Time Series Testing
04/28/2025

To assess the quality of a time series model, statistical testing is performed against a set of criteria. Based on the "scores" of the testing, different time series models can be ranked and the better ones are selected for further study.

Regression Testing

Regression tests are not only used for time series models but for linear models as well. Let a linear model be

$$y = \beta^T \mathbf{x} + \epsilon$$

where y is the response variable, $\mathbf{x}$ explanatory variable vector (dimension of m), β coefficient vector, and ϵ noise. Further, let $\hat{\beta}$ be the estimate vector for the coefficients based on n observations, the sum of squares of error (SSE) is a measure of the overall error terms of the model:

$$SSE = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \hat{\beta}_{ji} x_{ji} \right)^2$$

Often R^2 is used to test a linear model:

$$R^2 = \frac{SSE_0 - SSE}{SSE_0} \qquad \text{where } SSE_0 = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

and is also known as the adjusted total sum of squares. The test, R^2 , is called the coefficient of determination, since SEE is the error primarily from the noise and SSE_0 is the combined error of the model and the noise.

The mean squared error (MSE) is another measurement for the error term "normalized" by the number of observations minus the dimension of the coefficients (or degrees of freedom):

$$MSE = \frac{SSE}{n - (m+1)}$$

Using analysis of variance, it may determine that only a subset of the explanatory variables ($r < m$) are truly responsible for the response variable:

$$SSEr = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{r} \hat{\beta}_{ji} x_{ji} \right)^2$$

The F-statistic is defined as the ratio of the MSE of the reduced model (called MSR) and the MSE of the full model:

$$F = \frac{MSR}{MSE} = \frac{(SSEr - SSE)/(m-r)}{SSE/(n-m-1)}$$

If F is very small, the H_0 hypothesis should be rejected, *i.e.,* the reduced model is accepted.

However, often the value of r is not readily known and one may perform F-tests for several of r values. This approach is called stepwise multiple regression before selecting the best model. Likewise, another technique is used to perform ordinary regression using only k coefficients and compute the maximum likelihood estimator for the variance:

$$\hat{\sigma}_k^2 = \frac{SSE(k)}{n}$$

before selecting the best k for the smallest $\hat{\sigma}_k^2$. In practice, however, $\hat{\sigma}_k^2$ decreases monotonically as k increases. To compensate for this effect, Akaike proposed the following Information Criterion:

$$AIC = log(\hat{\sigma}_k^2) + \frac{n+2k}{n}$$

This is called Akaike's Information Criterion (AIC). Other researchers proposed additional modification on AIC, known as AIC, Bias Corrected (AICc) and Bayesian Information Criterion (BIC):

$$AICc = log(\hat{\sigma}_k^2) + \frac{n+k}{n-k-2}$$

$$BIC = log(\hat{\sigma}_k^2) + \frac{klog(n)}{n}$$

BIC is best suited for a large sample size whereas AICc is better for a smaller sample size.

Autocorrelation Testing

The residuals of a properly-fitted model approach to the white noise, *i.e.*, the autocorrelation of the residuals at lag h is sufficiently close to zero. Hypothesis testing of the autocorrelation is constructed for $H_0 : \rho_y(1) = \rho_y(2) = \cdots = \rho_y(h) = 0$, $H_1 : \exists i \in \{1, 2, \ldots, h\}\ \rho_y(i) \neq 0$.

Box-Pierce test statistic follows $\chi^2(h)$ approximately under the null hypothesis H_0:

$$Q_{B-P} = n \sum_{i=1}^{h} \hat{\rho}_y^2(i)$$

Ljung-Box test statistic follows $\chi^2(h)$ more precisely under H_0:

$$Q_{L-B} = n(n+2) \sum_{i=1}^{h} \frac{\hat{\rho}_y^2(i)}{n-i}$$

Given a linear model: $y_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p + r_t$, where r_t is the residual of the model and its estimate follows an autocorrelation model:

$$\hat{r}_t = \rho \hat{r}_{t-1} + \epsilon_t$$

where ρ is the autocorrelation coefficient and w_t white noise. The hypothesis for the test is constructed as $H_0 : \rho = 0, H_1 : \rho \neq 0$. Durbin-Watson test statistic is:

$$Q_{D-W} = \frac{\sum_{t=2}^{T} \left(\hat{r}_t - \hat{r}_{t-1}\right)^2}{\sum_{t=1}^{T} \hat{r}_t^2}$$

If the test statistic is close to zero, it suggests a strong, positive autocorrelation since $\hat{r}_t$ and $\hat{r}_{t-1}$ have similar values. Conversely, if the test statistic is large (maximum possible value of 4), it suggests a strong, negative autocorrelation because $\hat{r}_t$ and $\hat{r}_{t-1}$ have opposite signs. A non-autocorrelation condition would yield the test statistic close to 2.

Autocorrelation plots against the time lag, h, can be used as a diagnostic tool for models that best fit the data. Autocorrelation function (ACF) and partial autocorrelation function (PACF) are often used.

$$ACF(h) = \frac{\hat{K}(h)}{\hat{K}(0)}$$

where

$$\hat{K}(h) = \frac{1}{T} \sum_{t=1}^{T-h} \left(y_{t+h} - \overline{y}\right)\left(y_t - \overline{y}\right)$$

$$PACF(h) = corr\left(y_{t+h}, y_t \mid y_{t+h-1}, \ldots, y_{t+1}\right)$$

The table below shows the characteristics of ACF(h) and PAFC(h) plots for an oder p autoregression, AR(p), model, or an order q moving average, MA(q), model.

	AR(p)	**MA(q)**
ACF(h)	decreasing geometrically	zero for h > q
PACF(h)	zero for h > p	decreasing geometrically

<u>Unit Root Testing</u>

Testing for stationarity of a time series is referred as a unit root test. If the autoregressive term of a time series has a unit root ($\phi_i = 1$), the process is not stationary. Rather, either the process has a deterministic trend, or is a random walk. Before one can perform a time series analysis, these trends must be removed either by a difference operator or subtracting the trend using a linear regression.

The Dickey-Fuller test performs the unit root test for an AR(1) process:

$$y_t = \phi y_{t-1} + \epsilon_t$$

Applying the difference operator:

$$\nabla y_t = (\phi - 1)y_{t-1} + \epsilon_t$$

The null hypothesis ($H_0 : \phi = 1$) is that the process has a unit root and the alternative hypothesis ($H_1 : \phi \neq 1$) is that all the roots are outside the unit circle.

The augmented Dickey-Fuller tests apply this method to general AR(p) processes.

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t$$

$$\nabla y_t = (\phi_1 - 1)y_{t-1} + \sum_{i=1}^{p-1} (\phi_{i+1} - 1)\nabla y_{t-i} + \epsilon_t$$

An alternative to the augmented Dickey-Fuller test is the Phillips-Perron test which is designed such that the test is more robust to the conditions of the time series.