Time Series Regression (Ch.2)

We start from classical regression

$$x_t = \beta_0 + \beta_1 z_{t1} + \beta_2 z_{t2} + \dots + \beta_q z_{tq} + w_t = \vec{\beta}^T \vec{z}_t + w_t \quad -(1$$

where $\vec{\beta} = (\beta_0, \beta_1, \dots, \beta_q)$ are unknown fixed regression coefficients, and $w_t \sim N(0, \sigma_w^2)$ iid. The predictors $\vec{z}_t$ are assumed to be independent series

Rewrite (1) in matrix-vector form

$$\vec{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \vec{Z}\vec{\beta} + \vec{w} = \begin{pmatrix} \vec{z}_1^T \\ \vdots \\ \vec{z}_n^T \end{pmatrix} \vec{\beta} + \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}$$

The OLS estimator of $\vec{\beta}$ is

$$\hat{\beta} = \min_{\beta} \| \vec{X} - \vec{Z}\vec{\beta} \|^2 = \min_{\beta} \sum_{t=1}^{n} (x_t - \vec{z}_t^T \vec{\beta})^2 = (\vec{Z}^T \vec{Z})^{-1} \vec{Z}^T \vec{X}$$

Since $E(\hat{\vec{\beta}}) = E\left((\vec{Z}^T\vec{Z})^{-1} \vec{Z}^T (\vec{Z}\vec{\beta} + \vec{w}) \right) = E\left[\vec{\beta} + (\vec{Z}^T\vec{Z})^{-1}\vec{Z}^T \vec{w} \right] = \vec{\beta}$,
$\hat{\beta}$ is an unbiased estimators.

Given $\vec{Z}$, $\quad Var(\hat{\vec{\beta}}) = Var\left((\vec{Z}^T\vec{Z})^{-1} \vec{Z}^T \vec{w} \right) \qquad (\because Var(Aw) = A\, Var(w) A^T)$

$$= (\vec{Z}^T\vec{Z})^{-1} \vec{Z}^T \, Var(\vec{w}) \, \vec{Z} (\vec{Z}^T\vec{Z})^{-1}$$

$$= \sigma_w^2 (\vec{Z}^T \vec{Z})^{-1}$$

$\therefore$ If $\vec{w} \sim N(0, \sigma_w^2 I)$, $\quad \hat{\beta} \sim N(\vec{\beta}, \sigma_w^2 C)$, $\quad C = (\vec{Z}^T\vec{Z})^{-1} = \left(\sum_{t=1}^{n} \vec{z}_t \vec{z}_t^T \right)$

Let the sum of squares error (SSE) to be

$$SSE = \| \vec{X} - \vec{Z}\hat{\beta} \|^2 = \sum_{t=1}^{n} (\vec{x}_t - \vec{z}_t^T \hat{\beta})^2$$

An unbiased estimator for the variance σ_w^2 is

$$\underset{\text{(mean squared error)}}{S_w^2 = MSE} = \frac{SSE}{n-(q+1)}$$

The test statistic $\quad t_i = \dfrac{\hat{\beta}_i - \beta_i}{S_w \sqrt{C_{ii}}} \sim t_{n-(q+1)}$

where C_{ii} denotes the ith diagonal element of C.

If we want to test $H_0: \beta_{r+1} = \dots = \beta_q = 0$, i.e.

$$X_t = \beta_0 + \beta_1 z_{t1} + \dots + \beta_r z_{tr} + W_t \quad \text{—— (2)}$$

we have

$$F = \frac{(SSE_r - SSE)/(q-r)}{SSE/(n-q-1)} \sim F_{q-r,\, n-q-1} \text{ under } H_0$$

where $SSE_r = \sum_{t=1}^{n} (X_t - (\hat{\beta}_0^r + \hat{\beta}_1^r z_{t1} + \dots + \hat{\beta}_r^r z_{tr}))^2$ is the sum of squares error under H_0 and $\hat{\beta}_i^r$ is the corresponding OLS estimator.

In the case of $r=0$, $SSE_0 = \sum_{t=1}^{n} (X_t - \bar{X})^2$

the term $R^2 = \dfrac{SSE_0 - SSE}{SSE_0}$ is called the coefficient of determination.

Although R^2 can be used to measure the goodness of fit of a model, we can easily get $R^2 = 1$ for q, number of predictors, greater than n ($\because$ overfitting)

To avoid including too many irrelevant predictors in the model, various model selection methods have been developed.

For example, from a set of model candidates, choose the one with k unknown parameters to minimize

Definition 2.1 $\quad AIC = \log \hat{\sigma}_k^2 + \dfrac{n+2k}{n}, \quad \hat{\sigma}_k^2 = \dfrac{SSE(k)}{n}$

where $SSE(k)$ denotes the sum of squares error under the model with k regression coefficients.

When n is small (small sample size), it is proposed to use

Def. 2.2 $\quad AIC_c = \log \hat{\sigma}_k^2 + \dfrac{n+k}{n-k-2}$ (see Problems 2.4 and 2.5)

When we are only interested in significant predictors, then we may consider

Def. 2.3 $\quad BIC = \log \hat{\sigma}_k^2 + \dfrac{k \log n}{n}$

Example 2.2 | In this example, 4 models are considered

$$M_t = \beta_0 + \beta_1 t + \beta_2 (T_t - T_\cdot) + \beta_3 (T_t - T_\cdot)^2 + \beta_4 P_t + w_t$$

Model (i): $\beta_2 = \beta_3 = \beta_4 = 0$
(ii): $\beta_3 = \beta_4 = 0$
(iii): $\beta_4 = 0$
(iv): Full model

T_t: temperature
P_t: particulate level (pollut
$T_\cdot = \frac{1}{n} \sum_{t=1}^{n} T_t$
$n = 508$

Model	k	SSE	df (=n-k)	MSE	R^2	AIC	BIC
(i)	2	40,020	506	79.0	.21	5.38	5.40
(ii)	3	31,413	505	62.2	.38	5.14	5.17
(iii)	4	27,985	504	55.5	.45	5.03	5.07
(iv)	5	20,508	503	40.8	.60	4.72	4.77

It suggests the full model (iv) is the most suitable

To test $H_0: \beta_2 = \beta_3 = \beta_4 = 0$, we consider the test statistic

$$F = \frac{(SSE(2) - SSE(5))/(5-2)}{SSE(5)/(508-5)} = 160 > 5.51$$

which is the 0.999^{th} quantile of $F_{3,503}$. $\therefore H_0$ is rejected.

How to handle non-stationary time series? (Sect. 2.2)

For X_t non-stationary, we assume $X_t = \mu_t + y_t$, where μ_t is nonstationary and y_t is stationary.

By including the constant term in μ_t, we can assume $E y_t = 0$ and $Var(y_t) = \sigma_y^2$. It makes y_t look like an error term (although they are not iid)

$\therefore$ To handle X_t, one way is by "detrending"

1. Find a set of predictors that are relevant to X_t but nonstationary
2. Fit a linear regression model on X_t, e.g. $X_t = \beta_0 + \beta_1 t + y_t$ with unknown coefficients estimated by ordinary least squares.
3. Apply time series analysis on $\hat{y}_t = X_t - \hat{\beta}_0 - \hat{\beta}_1 t$

We have seen that a random walk process is non-stationary. Consider $\mu_t = \delta + \mu_{t-1} + w_t$, where $w_t \sim wn(0, \sigma_w^2)$ independent of y_t, then differencing x_t yields a stationary process

$$x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1})$$
$$= \delta + w_t + y_t - y_{t-1}$$

Recall that $E(y_t) = 0$, $Cov(y_{t+h}, y_t) = \gamma_y(h)$ does not depend on t

$$E(x_t - x_{t-1}) = \delta$$

$$Cov(x_{t+h} - x_{t+h-1}, x_t - x_{t-1}) = Cov(\delta + w_{t+h} + y_{t+h} - y_{t+h-1}, \delta + w_t + y_t - y_{t-1})$$
$$= Cov(w_{t+h}, w_t) + 2\gamma_y(h) - \gamma_y(h-1) - \gamma_y(h+1)$$

Define $\nabla x_t = x_t - x_{t-1} = x_t - Bx_t = (1-B)x_t$

where B is the backshift operator. For any time series y_t, $By_t = y_{t-1}$, $B^2 y_t = B(By_t) = By_{t-1} = y_{t-2}$. In general $B^k x_t = x_{t-k}$

Similarly, we can define the forward-shift operator B^{-1} so that

$$B^{-1} x_{t-1} = x_t$$

For nonstationary x_t, if $y_t = x_t - x_{t-1} = (1-B)x_t$ is still nonstationary, we can consider differencing y_t again to get

$$z_t = y_t - y_{t-1} = (1-B)y_t = (1-B)(1-B)x_t = (1-B)^2 x_t$$

Note that it is the same as

$$z_t = y_t - y_{t-1} = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) = x_t - 2x_{t-1} + x_{t-2}$$
$$= x_t - 2Bx_t + B^2 x_t$$
$$= (1 - 2B + B^2)x_t = (1-B)^2 x_t$$

Definition 2.5 Differences of order d are defined as

$$\nabla^d = (1-B)^d$$

It is possible that $\{x_t\}$ rejects the null hypothesis of stationarity but $\{f(x_t)\}$ accepts. One commonly used function is $f(x_t) = \log x_t$.

We have seen that the technique of moving average (e.g. $V_t = \frac{W_{t-1} + W_t + W_{t+1}}{3}$) can reduce the variance of the noise and hence we can have better estimates of parameters.

In general, consider $m_t = \sum_{j=-k}^{k} a_j X_{t-j}$, where $a_j \geq 0$ and $\sum a_j = 1$

$\boxed{\text{Example 2.12}}$ Kernel Smoothing

For observations $X_1, X_2, \ldots, X_n$, we construct

$$m_t = \sum_{i=1}^{n} w_i(t) X_i, \quad \text{where} \quad w_i(t) = \frac{K\left(\frac{t-i}{b}\right)}{\sum_{j=1}^{n} K\left(\frac{t-j}{b}\right)}$$

K is usually chosen to be symmetric and $K(x) = 0$ or very close to 0 if $|x|$ is large (e.g. $|x| > 1$). A common choice is the density of $N(0,1)$, i.e. $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$