LINEAR REGRESSION ANALYSIS

MODULE – XI

Lecture - 33

Autocorrelation

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

One of the basic assumptions in linear regression model is that the random error components or disturbances are identically and independently distributed. So in the model $y = X\beta + u,$ it is assumed that

$$E(u_t, u_{t-s}) = \begin{cases} \sigma_u^2 & \text{if } \; s = 0 \\ 0 & \text{if } \; s \neq 0, \end{cases}$$

i.e., the correlation between the successive disturbances is zero.

In this assumption, when $E(u_t, u_{t-s}) = \sigma_u^2,\ s = 0$ is violated, i.e., the variance of disturbance term does not remains constant, then **problem of heteroskedasticity** arises.

When $E(u_t, u_{t-s}) = 0,\ s \neq 0$ is violated, i.e., the variance of disturbance term remains constant though the successive disturbance terms are correlated, then such problem is termed as **problem of autocorrelation.**

When autocorrelation is present, some or all off diagonal elements in $E(uu')$ are nonzero.

Sometimes the study and explanatory variables have a natural sequence order over time, i.e., the data is collected with respect to time. Such data is termed as **time series data.** The disturbance terms in time series data are serially correlated.

The **autocovariance** at lag *s* is defined as

$$\gamma_s = E(u_t, u_{t-s});\ s = 0, \pm 1, \pm 2, \ldots .$$

At zero lag, we have constant variance, i.e.,

$$\gamma_0 = E(u_t^2) = \sigma^2.$$

The **autocorrelation coefficient** at lag *s* is defined as

$$\rho_s = \frac{E(u_t u_{t-s})}{\sqrt{Var(u_t)Var(u_{t-s})}};\ s = 0, \pm 1, \pm 2, \ldots$$

Assume ρ_s and γ_s are symmetrical in *s*, i.e., these coefficients are constant over time and depend only on length of lag *s*.

The autocorrelation between the successive terms (u_2 and u_1), (u_3 and u_2),…, (u_n and u_{n-1}) gives the autocorrelation of order one, i.e., ρ_1.

Similarly, the autocorrelation between the successive terms $(u_3$ and $u_1), (u_4$ and $u_2), \ldots, (u_n$ and $u_{n-2})$ gives the autocorrelation of order two, i.e., ρ_2.

Source of autocorrelation

Some of the possible reasons for the introduction of autocorrelation in the data are as follows:

1. Carryover of effect, atleast in part, is an important source of autocorrelation. For example, the monthly data on expenditure on household is influenced by the expenditure of preceding month. The autocorrelation is present in cross-section data as well as time series data. In the cross-section data, the neighboring units tend to be similar with respect to the characteristic under study. In time series data, the time is the factor that produces autocorrelation. Whenever some ordering of sampling units is present, the autocorrelation may arise.

2. Another source of autocorrelation is the effect of deletion of some variables. In regression modeling, it is not possible to include all the variables in the model. There can be various reasons for this, e.g., some variable may be qualitative, sometimes direct observations may not be available on the variable etc. The joint effect of such deleted variables gives rise to autocorrelation in the data.

3. The misspecification of the form of relationship can also introduce autocorrelation in the data. It is assumed that the form of relationship between study and explanatory variables is linear. If there are log or exponential terms present in the model so that the linearity of the model is questionable then this also gives rise to autocorrelation in the data.

4. The difference between the observed and true values of variable is called measurement error or errors–in-variable. The presence of measurement errors on the dependent variable may also introduce the autocorrelation in the data.

Structure of disturbance term

Consider the situation where the disturbances are autocorrelated,

$$E(\varepsilon\varepsilon') = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{bmatrix}$$

$$= \gamma_0 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{bmatrix}$$

$$= \sigma_u^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{bmatrix}.$$

Observe that now there are (n + k) parameters- $\beta_1, \beta_2, ..., \beta_k, \sigma_u^2, \rho_1, \rho_2, ..., \rho_{n-1}$. These ($n$ + k) parameters are to be estimated on the basis of available n observations. Since the number of parameters are more than the number of observations, so the situation is not good from the statistical point of view. In order to handle the situation, some special form and the structure of the disturbance term is needed to be assumed so that the number of parameters in the covariance matrix of disturbance term can be reduced.

The following structures are popular in autocorrelation:

1. Autoregressive (AR) process.
2. Moving average (MA) process.
3. Joint autoregressive moving average (ARMA) process.

Estimation under the first order autoregressive process

Consider a simple linear regression model

$$y_t = \beta_0 + \beta_1 X_t + u_t,\ t = 1, 2, ..., n.$$

Assume u_t 's follow a first order autoregressive scheme defined as

$$u_t = \rho u_{t-1} + \varepsilon_t$$

where

$$|\rho| < 1,\ E(\varepsilon_t) = 0,$$

$$E(\varepsilon_t, \varepsilon_{t+s}) = \begin{cases} \sigma_\varepsilon^2 & \text{if } s = 0 \\ 0 & \text{if } s \neq 0 \end{cases}$$

for all $t = 1, 2, ..., n$ where ρ is the first order autocorrelation between u_t and u_{t-1}, $t = 1, 2, ..., n.$ Now

$$\begin{aligned} u_t &= \rho u_{t-1} + \varepsilon_t \\ &= \rho(u_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \vdots \\ &= \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + ... \\ &= \sum_{r=0}^{\infty} \rho^r \varepsilon_{t-r} \end{aligned}$$

$$E(u_t) = 0$$

$$E(u_t^2) = E(\varepsilon_t^2) + \rho^2 E(\varepsilon_{t-1}^2) + \rho^4 E(\varepsilon_{t-2}^2) + ...$$

$$= (1 + \rho^2 + \rho^4 + ...)\sigma_\varepsilon^2 \quad (\varepsilon_t^{'}s \text{ are serially independent})$$

$$E(u_t^2) = \sigma_u^2 = \frac{\sigma_\varepsilon^2}{1-\rho^2} \text{ for all } t.$$

$$E(u_t u_{t-1}) = E\left[\left(\varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \ldots\right)\times\left(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \rho^2\varepsilon_{t-3} + \ldots\right)\right]$$

$$= E\left[\left\{\varepsilon_t + \rho\left(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \ldots\right)\right\}\left\{\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \ldots\right\}\right]$$

$$= \rho E\left[\left(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \ldots\right)^2\right]$$

$$= \rho\sigma_u^2.$$

Similarly,

$$E(u_t u_{t-2}) = \rho^2\sigma_u^2.$$

In general,

$$E(u_t u_{t-s}) = \rho^s\sigma_u^2$$

$$E(uu') = \Omega = \sigma_u^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}.$$

Note that the disurbance terms are no more independent and $E(uu') \neq \sigma^2 I.$ The disturbance are nonspherical.

Consequences of autocorrelated disturbances

Consider the model with first order autoregressive disturbances

$$\underset{n\times 1}{y} = \underset{n\times k}{X}\,\underset{k\times 1}{\beta} + \underset{n\times 1}{u}$$

$$u_t = \rho u_{t-1} + \varepsilon_t,\ t = 1,2,\ldots,n$$

with assumptions

$$E(u) = 0,\ E(uu') = \Omega$$

$$E(\varepsilon_t) = 0,\ E(\varepsilon_t\varepsilon_{t+s}) = \begin{cases} \sigma_\varepsilon^2 & \text{if } s = 0 \\ 0 & \text{if } s \neq 0 \end{cases}$$

where Ω is a positive definite matrix.

The ordinary least squares estimator of β is

$$\begin{aligned} b &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}X'(X\beta + u) \end{aligned}$$

$$b - \beta = (X'X)^{-1}X'u$$

$$E(b - \beta) = 0.$$

So OLSE remains unbiased under autocorrelated disturbances.

The covariance matrix of b is

$$\begin{aligned} V(b) &= E(b-\beta)(b-\beta)' \\ &= (X'X)^{-1}X'E(uu')X(X'X)^{-1} \\ &= (X'X)^{-1}X'\Omega X(X'X)^{-1} \\ &\neq \sigma_u^2(X'X)^{-1}. \end{aligned}$$

The residual vector is

$$e = y - Xb = \bar{H}y = \bar{H}u$$

$$e'e = y'\bar{H}y = u'\bar{H}u$$

$$E(e'e) = E(u'u) - E\left[u'X(X'X)^{-1}X'u\right]$$

$$= n\sigma_u^2 - tr(X'X)^{-1}X'\Omega X.$$

Since $s^2 = \dfrac{e'e}{n-1}$, so

$$E(s^2) = \frac{\sigma_u^2}{n-1} - \frac{1}{n-1}tr(X'X)^{-1}X'\Omega X,$$

so s^2 is a biased estimator of σ^2. In fact, s^2 has downward bias.

Application of OLS fails in case of autocorrelation in the data and leads to serious consequences as

- overly optimistic view from R^2.
- narrow confidence interval.
- usual t-ratio and F - ratio tests provide misleading results.
- prediction may have large variances.

Since disturbances are nonspherical, so generalized least squares estimate of β yields more efficient estimates than OLSE.

The GLSE of β is

$$\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

$$E(\hat{\beta}) = \beta$$

$$V(\hat{\beta}) = \sigma_u^2(X'\Omega^{-1}X)^{-1}.$$

The GLSE is best linear unbiased estimator of β.