Regression Analysis

Chapter 8: The Problem of Correlated Errors

Kyusang Yu and Sunghwan Kim

Department of Applied Statistics

Konkuk University

Spring 2019

Autocorrelation

- One of the standard assumption in the regression model is that the error terms ϵ_i and ϵ_j, associated with the i-th and j-th observations, are uncorrelated.
- When the observations have a natural sequential order, the correlation is referred to as *autocorrelation*.
  - Ex 1. Successive residuals in economic time series tend to be positively correlated.
  - Ex 2. Observations sampled from adjacent experimental plots or areas tend to have residuals that are correlated.

Autocorrelation

- Reasons for occurrence
  - Temporal dimension
  - Spatial dimension
  - Omission of a correlated predictor

Effects of autocorrelation

1. Least squares estimates of the regression coefficients are unbiased but are not efficient in the sense that they no longer have minimum variance.
2. The estimate of σ^2 and the standard errors of the regression coefficients may be seriously understated (may give a spurious impression of accuracy).
3. The confidence intervals and the various tests of significance commonly employed would no longer be strictly valid.

Types of autocorrelation

- We distinguish between two types of autocorrelation and describe methods for dealing with each.
  1. The first type is only autocorrelation in appearance because of omission of a autocorrelated predictor (can be resolved once we add that variable).
  2. The second type of autocorrelation may be referred to as pure autocorrelation (need a transformation of the data to remove such autocorrelation).

Consumer Expenditure and Money Stock (CEMS)

- See data in Table 8.1 (p. 211)
- Y: consumer expenditure from 1952 to 1956
- X: the stock of money
- A simplified version of the quantity theory of money suggests a model given by

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t,$$

  where β_0 and β_1 are constants, ϵ_t the error term.
- Economists are interested in estimating β_1 (multiplier) and its standard error.

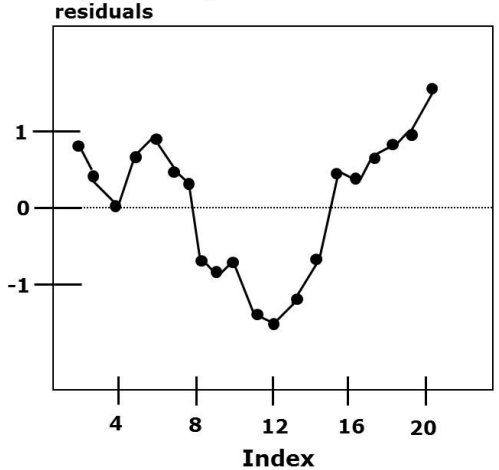
CEMS - continued

- See the regression results in Table 8.2 (p. 211).
- $\hat{\beta}_1 = 2.30$ with $p - \text{value} < 0.0001$
- $R^2 = 0.957$
- The analysis is complete if the basic assumptions were valid.
- If there are indications that autocorrelation is present, the model should be re-estimated after eliminating the autocorrelation.

CEMS - continued

- From Figure 8.1 (p. 212), which is referred to as the index plot, we see that the first seven residuals are positive, the next seven negative, and the last six positive.
- This pattern suggest that the error terms in the model are correlated and some additional analysis is required.

Fig. 8.1 **index plot of the standardized residuals**



Run test

- A run of successes or failures is a series of successes or failures (i.e., run = sequence or series).
- For example, in Figure 8.1

$$+++++++---------++++++ \Rightarrow \text{three runs}$$

- Notations
  - μ = the expected number of runs
  - σ^2 = the variance of the number of runs
  - n_1 = the number of positive values
  - n_2 = the number of negative values

Run test

- H_0 : random (no autocorrelation)
- Test statistic:

$$Z = \frac{r - \mu}{\hat{\sigma}} \approx N(0, 1),$$

where

$$\begin{aligned} r &= \text{run}, \\ \mu &= \frac{2n_1 n_2}{n_1 + n_2} + 1, \\ \sigma^2 &= \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}. \end{aligned}$$

Run test for CEMS data

- Observed values: $n_1 = 13$, $n_2 = 7$, $r = 3$
- Under H_0 : random (no correlation), $\mu = 10.1$, $\sigma = 1.97$

$$Z = \frac{3 - 10.1}{1.97} = -3.6$$

- The deviation of 3.6 from the expected number of runs is more than twice the standard deviation, indicating a significant departure from randomness.

Durbin-Watson statistic

- The Durbin-Watson statistic is the basis of a popular test of autocorrelation in regression analysis.
- The test is based on the assumption that successive error s are correlated, namely,

$$\epsilon_t = \rho\epsilon_{t-1} + \omega_t, \quad |\rho| < 1,$$

  where ρ is the correlation coefficients between ϵ_t and ϵ_{t-1} , and ω_t is normally independently distributed with zero mean and constant variance.
- In this case, the errors are said to have *first-order autoregressive structure* or *first order autocorrelation.*

Durbin-Watson statistic - continued

- Note that when $\rho = 0$, the ϵ's are uncorrelated.
- Test $H_0: \quad \rho = 0 \quad vs. \quad H_1: \quad \rho > 0$
- Since ρ is unknown, estimate the parameter ρ by $\hat{\rho}$, where

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2},$$

  where e_i is the i-th OLS residual.
- The Durbin-Watson statistic is defined as

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}.$$

Durbin-Watson statistic - continued

- An approximate relationship between d and $\hat{\rho}$ is

$$d \approx 2(1 - \hat{\rho})$$

- Note that d has a range of 0 to 4.
- d is close to 2 when $\rho = 0$ and near to zero when $\rho = 1$.
- The closer the sample value of d to 2, the firmer the evidence that there is no autocorrelation present in the error.

Durbin-Watson statistic - continued

- The test for positive autocorrelation:
  1. If $d < d_L$, reject H_0.
  2. If $d > d_U$, do not reject H_0.
  3. If $d_L < d < d_U$, the test is inconclusive.
- The values of (d_L, d_U) for different percentage points have been tabulated by Durbin and Watson (1951).
- A table is provided in the Appendix at the end of the book (Tables A.6 and A.7).

Example: CEMS data

- $d = 0.328$
- With $n = 20$, $p = 1$, under $\alpha = 0.05$, $d_L = 1.20$ and $d_U = 1.41$ (Table A.6, p.378).
- Under the significance level of 0.05, H_0 is rejected, showing that autocorrelation is present.

Example: CEMS data

- If $d > d_U$, no further analysis is needed.
- If $d < d_L$, following approaches may be followed:
  1. Work with transformed variables, or
  2. Introduce additional variables that have time-ordered effects.
- When $d_L < d < d_U$, reestimate the equation using the above approaches to see if any major changes occur.

Removal of autocorrelation by transformation

- One method for adjusting the model is the use of a transformation that involves the unknown autocorrelation parameter, ρ.
- From model $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$, ϵ_t and ϵ_{t-1} can expressed as

$$\begin{aligned} \epsilon_t &= y_t - \beta_0 - \beta_1 x_t \\ \epsilon_{t-1} &= y_{t-1} - \beta_0 - \beta_1 x_{t-1}. \end{aligned}$$

- Substituting these in $\epsilon_t = \rho\epsilon_{t-1} + \omega_t$, and rearranging term, we obtain

$$\begin{aligned} y_t - \rho y_{t-1} &= \beta_0(1-\rho) + \beta_1(x_t - \rho x_{t-1}) + \omega_t, \\ y_t^* &= \beta_0^* + \beta_1^* x_t^* + \omega_t. \end{aligned}$$

Removal of autocorrelation by transformation

- Since the ω's are uncorrelated, this model is a linear model with uncorrelated errors.
- The estimates of the parameters in the original equations are

$$\hat{\beta}_0 = \frac{\hat{\beta}_0^*}{1-\hat{\rho}} \quad \text{and} \quad \hat{\beta}_1 = \hat{\beta}_1^*.$$

Removal of autocorrelation by transformation

- The value of ρ is unknown and has to be estimated from the data using iterative procedure:
  1. Compute the OLS estimates of β_0 and β_1 by fitting the model, $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$, to the data.
  2. Calculate the residuals and estimate ρ using

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}.$$

Removal of autocorrelation by transformation

3. Fit the equation $y_t^* = \beta_0^* + \beta_1^* X_t^* + \omega_t$ using the variables $y_t - \hat{\rho} y_{t-1}$ and $x_t - \hat{\rho} x_{t-1}$ as a response and predictor variables, respectively, and obtain

$$\hat{\beta}_0 = \frac{\hat{\beta}_0^*}{1 - \hat{\rho}} \quad \text{and} \quad \hat{\beta}_1 = \hat{\beta}_1^*.$$

Removal of autocorrelation by transformation

4. Examine the residuals of the newly fitted equation. If the new residuals continue to show autocorrelation, repeat the entire procedure using the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ as estimates of β_0 and β_1 instead of the original least squares estimates. On the other hand, if the new residuals show no autocorrelation, the procedure is terminated and the fitted equation for the original data is

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t.$$

Example - CEMS data

- $d = 0.328$ - highly significant!
- $\hat{\rho} = 0.751$
- On fitting the regression equation to the variables $(y_t - 0.751 y_{t-1})$ and $(X_t - 0.751 X_{t-1})$, we have $d = 1.43$.
- Since with $n = 19$, $p = 1$, under $\alpha = 0.05$, $d_L = 1.18$ and $d_U = 1.40$, $H_0 : \rho = 0$ is not rejected.

Example - CEMS data

- Since the fitted equation is $\hat{y}_t^* = -53.64 + 2.64x_t^*$, the fitted equation in terms of the original variables is

$$\hat{y}_t = -215.31 + 2.64x_t.$$

- Note that OLS equation was $\hat{y}_t = -154.7 + 2.3x_t$.
- See Figure 8.2 (p. 216).

Autocorrelation and Missing Variables

- Autocorrelation
  - Index plot
  - Runs test
  - DW statistic
- Source of autocorrelation
  - Missing variables (more useful if one can understand the autocorrelation by missing variables)
  - Pure autocorrelation of errors (an action of last resort)

Analysis of Housing Starts

- Project by a midwestern construction industry association
- Objective: understand the relationship between housing starts and population growth.
- Approach
  - To develop annual data on regional housing starts
  - To relate these data to potential home buyers

Analysis of Housing Starts

- Measurements
  - H = housing starts (Housing starts refers to the number of privately owned new homes (technically housing units) on which construction has been started in a given period.)
  - P = population size
  - D = availability for mortgage money index
- Tentative model
  - $H_t = \beta_0 + \beta_1 P_t + \epsilon_t$
  - $R^2 = .925$
  - Estimate of $\beta_1 = .0714$
  - Index plot (Figure 8.3, p. 220) and DW-statistic ($d = 0.621$) suggests a strong autocorrelation.

Analysis of Housing Starts - continued

- A method: add the availability of mortgage money (see Table 8.6, p. 221)
  - $H_t = \beta_0 + \beta_1 P_t + \beta_2 D_t + \epsilon_t$
  - $d = 1.85$ does not indicate autocorrelation.
  - Index plot (Figure 8.4, p. 220) has improved compared with Figure 8.3.
  - D is more important than P (c.f., t-test).

Implications

1. A large value of R^2 does not necessarily imply that the data have been fitted and explained well.
2. DW-statistic and the index plot shows autocorrelation but do not suggest the source of it.
   - In general, a significant value of the DW statistic should be interpreted as an indication that a problem exists.
   - Pure autocorrelation of the errors (Money Stock data)
   - Omission of other variables (Housing Starts data)

Remark

- Suppose the model $H_t = \beta_0 + \beta_1 P_t + \beta_2 D_t + \epsilon_t$ is correct, where ϵ_t is uncorrelated.
- Consider the model $H_t = \tilde{\beta}_0 + \tilde{\beta}_1 P_t + \tilde{\epsilon}_t$, then

$$\begin{aligned} \tilde{\epsilon}_t &= H_t - (\tilde{\beta}_0 + \tilde{\beta}_1 P_t) \\ &= \beta_0 + \beta_1 P_t + \beta_2 D_t + \epsilon_t - (\tilde{\beta}_0 + \tilde{\beta}_1 P_t) \\ &= (\beta_0 - \tilde{\beta}_0) + (\beta_1 - \tilde{\beta}_1) P_t + \beta_2 D_t + \epsilon_t \end{aligned}$$

1. Suppose $\beta_0 - \tilde{\beta}_0 = \beta_1 - \tilde{\beta}_1 = 0$. If D_t is autocorrelated, then $\tilde{\epsilon}_t$ is also autocorrelated.
2. Suppose $(\beta_1 - \tilde{\beta}_1) P_t + \beta_2 D_t$ is autocorrelated, then $\tilde{\epsilon}_t$ is also autocorrelated.
3. Hence, $\tilde{\epsilon}_t$ can be autocorrelated even if ϵ_t is uncorrelated.

Limitations of Durbin-Watson Statistic

- Suppose that the pattern of time dependences other than first order exist.
- Index plot is still informative.
- DW may not yield much valuable information.
- DW is designed for first order autocorrelation but not for higher-order time dependence.

Ski sales data

- Objective: Examine a relationship of quarterly sales of ski equipment (S) to personal disposable income (PDI).
- The model: $S_t = \beta_0 + \beta_1 PDI_t + \epsilon_t$
- Table 8.7 (p. 222) appears to be encouraging.
- $d = 1.968$ does not indicate autocorrelation.
- Index plot shows a time dependence of the residuals.
- A seasonal effect (summer v.s. winter) seems to overlooked.

Indicator variables to remove seasonality

- Example: Ski sales data
  - The pattern of residuals suggests that there is a seasonal effect.
  - This seasonal effect can be characterized by defining an indicator (dummy) variable that takes the value 1 for each winter quarter (Q1 and Q4) and is set equal to zero for each summer quarter (Q2 and Q3).
  - See the expanded data with a seasonal effect (Table 8.8, p. 224).

Indicator variables to remove seasonality

- The expanded model is

$$S_t = \beta_0 + \beta_1 PDI_t + \beta_2 Z_t + \epsilon_t,$$

  where Z_t is the zero-one indicator variable for a seasonal effect.

- The expanded model can be represented by the two models:
  - Winter season ($Z_t = 1$): $S_t = (\beta_0 + \beta_2) + \beta_1 PDI_t + \epsilon_t$
  - Summer season ($Z_t = 0$): $S_t = \beta_0 + \beta_1 PDI_t + \epsilon_t$
  - Parallel lines (Figure 8.6, p. 225)
  - Index plot does not show a seasonal pattern.
  - The precision of the estimated marginal effect of PDI increased.

Indicator variables to remove seasonality

- DW statistic is only sensitive for first-order autocorrelaton.
  - 1st order $= -0.001$
  - 2nd, 4th, 6th, 8th-order correlation $= -.81, 0.76, -0.71, 0.73$
- When autocorrelation is indicated, the model should be refitted.
  - Inclusion of a missing variable
  - Transformation
- DW statistic has meaning when there is an ordering.
  - Time index
  - Alphabetic listing