Distributed Lag Model in Time Series Analysis

1 Objectives

The objectives of this lab are to:

- Understand the concept of the Distributed Lag Model (DLM) and its application in time series analysis.
- Learn how to model the dynamic relationship between the dependent and independent variables using lagged values.
- Estimate the parameters of the Distributed Lag Model using ordinary least squares (OLS).
- Interpret the effects of past values of the independent variable on the dependent variable.
- Analyze the significance of lag length in the model.

Prerequisites

Before starting the lab, students should have a basic understanding of the following concepts:

- Understand time series data, including stationarity and autocorrelation.
- **Linear Regression**
  - Understand Ordinary Least Squares (OLS) estimation.
  - Be able to interpret regression coefficients and evaluate model fit.
- Familiarity with libraries such as `statsmodels, pandas, matplotlib and numpy` to perform regression analysis and handle time series data
- **Model Evaluation Metrics**
  - Understand R-Squared, p-values, and residual analysis.
  - Assess the goodness-of-fit and the significance of model coefficients.

2 Theory

The Distributed Lag Model(DLM) is used to model the impact of past values of an independent variable (X_t) on the current value of the dependent variable (Y_t).
The general form of the Distributed Lag Model can be written as:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 X_{t-2} + \cdots + \beta_p X_{t-p} + \epsilon_t$$

Where:

- Y_t is the dependent variable at time t.

- X_t is the independent variable at time t.
- β_0 is the intercept term.
- $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients that capture the effect of the lagged values of the independent variable on the dependent variable.
- ϵ_t is the error term.
- p represents the number of lags included in the model.

The lag length p determines how many past values of the independent variable are considered in the model. The choice of p depends on the theoretical understanding of the problem and model selection criteria like the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC).

Model Estimation

The model is typically estimated using Ordinary Least Squares (OLS), which minimizes the sum of squared residuals:

$$\text{Minimize} \sum_{t=1}^{T} (Y_t - \hat{Y}_t)^2$$

Where $\hat{Y}_t$ is the predicted value of Y_t based on the estimated coefficients.

Evaluate the Goodness-of-Fit

The goodness of fit measures how well the model explains the variation in the dependent variable (Y_t).

- **R-Squared (R^2)**: This statistic tells us the proportion of variance in the dependent variable explained by the independent variables (and their lags) in the model. A higher R^2 suggests a better fit.

  $$R^2 = 1 - \frac{\sum_{t=1}^{T} (Y_t - \hat{Y}_t)^2}{\sum_{t=1}^{T} (Y_t - \bar{Y})^2}$$

  where:

  – Y_t is the actual value of the dependent variable (temperature),

  – $\hat{Y}_t$ is the predicted value from the model,

  – $\bar{Y}$ is the mean of the dependent variable.

- **Adjusted R-Squared**: This accounts for the number of predictors in the model and is often preferred to avoid overfitting when using multiple lagged variables.

Significance of the Coefficients

Each coefficient in the model (e.g., β_1, β_2, β_3) needs to be tested for statistical significance.

- **t-Tests**: For each coefficient, perform a t-test to determine if it is significantly different from zero (i.e., if the lagged temperature variables have a statistically significant impact on the current temperature).

  The t-statistic for each coefficient is calculated as:

  $$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

  where:

  - $\hat{\beta}_i$ is the estimated coefficient,
  - $SE(\hat{\beta}_i)$ is the standard error of the coefficient.

  If the absolute value of the t-statistic is large and the p-value is less than the chosen significance level (e.g., 0.05), the corresponding coefficient is considered statistically significant.

Overall Model Significance

The **F-Test** can be used to assess whether the model as a whole is statistically significant. The null hypothesis for the F-test is that none of the independent variables (including lagged ones) have any effect on the dependent variable.
The F-statistic is calculated as:

$$F = \frac{(\text{Explained Sum of Squares})/p}{(\text{Residual Sum of Squares})/(T - p - 1)}$$

where:

- p is the number of predictors (including lagged values),
- T is the total number of observations.

If the p-value from the F-test is less than the significance level (usually 0.05), you reject the null hypothesis and conclude that the model is significant as a whole.

Residual Analysis

After fitting the model, check the residuals (the difference between observed and predicted values) for any patterns. Ideally, residuals should be randomly distributed around zero. If there are patterns, this may suggest that the model is misspecified.

- **Homoscedasticity**: The residuals should have constant variance across all levels of the independent variables.
- **Autocorrelation**: Residuals should not be correlated. You can use the Durbin-Watson test to check for autocorrelation.

- **Goodness of Fit**: If the R^2 and adjusted R^2 values are high, this suggests the model fits the data well. If they are low, the model might need improvement.
- **Significance of Coefficients**: If the t-tests show that most or all of the coefficients are significant, the lagged temperatures have a meaningful effect on the current temperature.
- **Overall Model Significance**: If the F-test indicates the model is significant, this reinforces the notion that the lagged variables contribute to explaining the current temperature.
- **Residual Analysis**: If the residuals are well-behaved (i.e., random, homoscedastic, and no autocorrelation), this supports the validity of the model.

3 Tasks

The following tasks should be completed:

1. Estimate the parameters of a Distributed Lag Model using time series data for the following.

Date	**1-Dec**	**2-Dec**	**3-Dec**	**4-Dec**	**5-Dec**	**6-Dec**	**7-Dec**
Temperature (°C)	20	22.5	21.2	23.0	19.8	18.5	20.1

   - Plot the 7 days temperature data.

2. Analyze the impact of different lag lengths on the model.
3. Interpret the coefficients of the lagged variables.
4. Evaluate the model's fit and significance of the coefficients using statistical tests.
5. Perform Residual Analysis and overall model significance.
6. Visualize the model results with plots to show the relationship between Y_t and the lagged values of X_t.
7. Test the robustness of the model by varying the number of lags and checking the stability of the coefficients.

4 Expected Outcomes

By the end of this lab, students should be able to:

- Understand the concept and structure of the Distributed Lag Model.
- Estimate the parameters of the model using real data and interpret the results.
- Analyze the relationship between a dependent variable and its lagged independent variables.

- Perform model diagnostics and evaluate the fit of the model.
- Understand the impact of lag length on model estimation and make informed decisions on selecting an appropriate number of lags.

5 Assessment

Students will be assessed on the following:

- **Accuracy of Model Estimation (30%):** Correctly estimate the parameters and interpret the coefficients.
- **Model Interpretation (30%):** Properly interpret the relationship between the dependent and lagged independent variables.
- **Statistical Evaluation (20%):** Evaluate the model fit and significance of the coefficients using appropriate statistical tests.
- **Visualization (10%):** Provide clear and informative plots of the model results.
- **Report (10%):** A well-organized report summarizing the model, its interpretation, and findings.