✏️

Math behind OLS

Created	@September 20, 2024 12:32 AM
Class	Introduction to Regression with statsmodels in Python

Mathematics of Ordinary Least Squares (OLS)

Key Details

OLS is a method used in regression analysis to estimate the parameters of a linear relationship between two variables. The goal is to find the line (or plane in the case of multiple variables) that minimizes the sum of the squared differences between the observed values and the values predicted by the linear model.

General Form of Linear Regression

The general equation for a simple linear regression model is:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Where:

- y is the dependent variable.
- x is the independent variable.
- β_0 is the intercept.
- β_1 is the slope (coefficient) of the independent variable.
- ϵ is the error term (difference between observed and predicted values).

In this equation, OLS is used to find the values of β_0 and β_1 that minimize the sum of squared errors.

Sum of Squared Errors (SSE)

To minimize the error, we define the **Sum of Squared Errors (SSE)** as:

$SSE = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$

Where:

- y_i is the observed value of the dependent variable.
- x_i is the observed value of the independent variable.
- β_0 and β_1 are the regression parameters.
- n is the number of observations.

Minimizing SSE: Deriving the OLS Estimates

To minimize the SSE, we take the partial derivatives of the SSE with respect to β_0 and β_1, and set them equal to zero:

1. **For β_0 :**

   $\frac{\partial SSE}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$

2. **For β_1:**

   $\frac{\partial SSE}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$

Solving these two equations simultaneously gives us the OLS estimates for β_0 and β_1.

OLS Estimate for β_1

The solution for the slope β_1 (the coefficient of the independent variable) is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where:

- $\bar{x}$ is the mean of the independent variable x.
- $\bar{y}$ is the mean of the dependent variable y.

OLS Estimate for β_0

Once $\hat{\beta}_1$ is known, we can calculate the intercept β_0:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Coding Example: OLS in Python

Below is an example using Python to calculate the OLS regression line:

```
import numpy as np
import matplotlib.pyplot as plt

# Sample data (x: independent variable, y: dependent variable)
x = np.array([1, 2, 3, 4, 5])
y = np.array([2, 4, 5, 4, 5])

# Calculate means
x_mean = np.mean(x)
y_mean = np.mean(y)

# Calculate OLS coefficients
beta1 = np.sum((x - x_mean) * (y - y_mean)) / np.sum((x - x_m
ean) ** 2)
beta0 = y_mean - beta1 * x_mean

# Display results
print(f"Intercept (beta0): {beta0}")
print(f"Slope (beta1): {beta1}")

# Plot data and regression line
plt.scatter(x, y, label='Data')
plt.plot(x, beta0 + beta1 * x, color='red', label='OLS Line')
plt.xlabel('x')
plt.ylabel('y')
plt.legend()
plt.show()
```

Matrix/Vector Formulation of OLS

In multiple regression (with more than one independent variable), OLS can be expressed in matrix form. For simplicity, consider a regression with multiple independent variables:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Where:

- $\mathbf{y}$ is an $n \times 1$ vector of the dependent variable.
- $\mathbf{X}$ is an $n \times p$ matrix of independent variables (with each column representing one independent variable and each row representing an observation).
- $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients.
- $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of error terms.

OLS Estimate in Matrix Form

The OLS estimate for $\boldsymbol{\beta}$ in matrix form is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

This expression provides the vector of OLS estimates for the regression coefficients.

Key Takeaways

- OLS minimizes the sum of squared errors to estimate the regression coefficients.
- The regression line is the line that best fits the data, minimizing the difference between observed and predicted values.
- In multiple regression, OLS can be generalized using matrix algebra to handle multiple independent variables.