Linear Regression

by Juan Suazo Verger

1 Introduction to Linear Regression

Linear regression is a method used to model the relationship between a dependent variable y and one or more independent variables x. The simplest form, known as simple linear regression, assumes a linear relationship between two variables, represented by the equation:

$$\hat{y} = b_0 + b_1 x$$

where:

- $\hat{y}$ is the predicted value.
- b_0 is the intercept (constant term).
- b_1 is the slope of the regression line.
- x is the independent variable.

In the case of multiple linear regression, we extend this to multiple predictors $x_1, x_2, \ldots, x_n$:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

2 Types of Linear Regression and When to Use Them

2.1 Simple Linear Regression

Used when there is only one independent variable.

- **Example:** Predicting a student's test score based on the number of hours studied.

2.2 Multiple Linear Regression

Used when there are multiple independent variables.

- **Example:** Predicting a house price based on features such as size, number of bedrooms, and location.

2.3 Polynomial Regression

Used when the relationship between the independent and dependent variable is non-linear.

- **Example:** Predicting the growth of a plant over time, which may not follow a straight line.

2.4 Ridge Regression

Used when there is multicollinearity in the data, adding a penalty term to the OLS loss function.

- **Example:** Predicting stock prices with correlated predictors, like various economic indicators.

2.5 Lasso Regression

Similar to ridge regression but performs variable selection by penalizing the absolute size of coefficients.

- **Example:** Selecting the most relevant features in high-dimensional datasets, such as genetic data.

3 Using Categorical Variables in Linear Regression

Categorical variables can be included in linear regression models using dummy coding. This technique involves creating binary variables for each category of the categorical variable.

3.1 Example of Dummy Coding

Suppose we have a categorical variable ”Color” with three categories: Red, Blue, and Green. We would create two dummy variables:

- $D_1 = 1$ if the color is Red, 0 otherwise.
- $D_2 = 1$ if the color is Blue, 0 otherwise.

The Green category serves as the reference category. The regression equation would then look like:

$$\hat{y} = b_0 + b_1 D_1 + b_2 D_2 + b_3 x$$

where x represents other numerical predictors.

4 Ordinary Least Squares (OLS) Method

The Ordinary Least Squares (OLS) method estimates the coefficients of the linear regression model by minimizing the sum of the squared differences between the observed and predicted values:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

The OLS estimates can be obtained using matrix algebra:

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

where X is the matrix of independent variables, y is the vector of dependent variable observations, and $\hat{\beta}$ contains the estimated coefficients.

4.1 Assumptions of OLS

To make valuable use of the OLS method, several key assumptions must be met:

1. **Linearity:** The relationship between the independent and dependent variables must be linear. This can be assessed using scatterplots or residual plots.

2. **Independence:** Observations must be independent of each other. This assumption is critical in time series data where autocorrelation can occur.

3. **Homoscedasticity:** The variance of the error terms should be constant across all levels of the independent variables. Plotting residuals against predicted values can help check for homoscedasticity.

4. **No Autocorrelation:** In time series data, the residuals should not be correlated with each other. This can be tested using the Durbin-Watson statistic.

5. **No Perfect Multicollinearity:** The independent variables should not be perfectly correlated. Variance Inflation Factor (VIF) can be calculated to detect multicollinearity.

6. **Normality of Errors:** The residuals should be normally distributed, which can be checked using a Q-Q plot or a histogram of residuals.

If these assumptions are violated, the estimates produced by OLS may be biased or inefficient. In such cases, alternative methods or adjustments may be necessary.

5 Python Methods for Linear Regression

In Python, we can perform linear regression using various libraries. Some of the most common methods are:

5.1 Using `scikit-learn`

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```

This method fits a linear model to the data by minimizing the OLS loss function.

5.2 Using `statsmodels`

```
import statsmodels.api as sm
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
predictions = model.predict(X)
```

This method provides additional statistical information such as p-values and confidence intervals.

5.3 Using `numpy`

```
import numpy as np
beta = np.linalg.inv(X.T @ X) @ X.T @ y
```

This approach manually computes the OLS estimates using matrix algebra.

6 Other Regression Methods

Apart from OLS, there are other methods to estimate the parameters of a regression model, which are useful in different scenarios:

6.1 Generalized Least Squares (GLS)

GLS is used when there is heteroscedasticity (non-constant variance of errors) or autocorrelation in the data. The equation remains the same, but the method accounts for the structure of the error term.

6.2 Maximum Likelihood Estimation (MLE)

MLE maximizes the likelihood function, which represents the probability of observing the given data given the model parameters.

6.3 Bayesian Regression

In Bayesian regression, we assume prior distributions for the parameters and update them based on the observed data.

6.4 Kernel Regression

Kernel regression is a non-parametric technique that does not assume a specific form for the relationship between the variables.

6.5 Gaussian Process Regression

Gaussian Process Regression is a flexible probabilistic model that captures uncertainty about the regression function.

7 Key Metrics in Linear Regression

- **Coefficient of Determination** R^2: Measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- **P-value of t-statistic**: Determines whether the individual predictor variables are statistically significant.
- **F-statistic**: Assesses the overall significance of the regression model.

8 Conclusion

Linear regression is a fundamental tool in data science, offering a straightforward way to model and predict relationships between variables. While OLS is a common estimation method, understanding its assumptions and alternative methods such as GLS, MLE, and Bayesian regression can provide better results depending on the data structure.