Regression Analysis Primer

Your Name

July 13, 2024

1 Introduction

This primer provides an introduction to regression analysis, focusing on linear regression and its various components and steps. Regression analysis is a statistical method for examining the relationships between dependent and independent variables, commonly used for forecasting and predicting trends.

2 Linear Regression

Linear regression is one of the simplest and most widely used forms of regression analysis. It models the relationship between a dependent variable y and one or more independent variables x.

2.1 Model Representation

In simple linear regression, the relationship between the dependent variable y and a single independent variable x is represented by the following equation:

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

where:

- y is the dependent variable.
- x is the independent variable.
- β_0 is the y-intercept (constant term).
- β_1 is the slope of the regression line.
- ϵ is the error term (residual).

2.2 Assumptions

Linear regression analysis is based on several key assumptions:

1. **Linearity**: The relationship between the dependent and independent variables is linear.
2. **Independence**: Observations are independent of each other.
3. **Homoscedasticity**: The variance of the error terms is constant across all levels of the independent variable.
4. **Normality**: The error terms are normally distributed.

2.3 Estimation of Parameters

The parameters β_0 and β_1 are estimated using the least squares method, which minimizes the sum of the squared differences between the observed and predicted values of the dependent variable. The estimates are given by:

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

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

where $\bar{x}$ and $\bar{y}$ are the means of the independent and dependent variables, respectively.

2.4 Goodness of Fit

The goodness of fit of the regression model is typically measured by the coefficient of determination (R^2), which represents the proportion of the variance in the dependent variable that is predictable from the independent variable. It is calculated as:

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

where $\hat{y}_i$ is the predicted value of y_i.

2.5 Hypothesis Testing

Hypothesis tests are used to determine whether the estimated coefficients are statistically significant. The null hypothesis for the slope (β_1) is:

$$H_0 : \beta_1 = 0$$

which means there is no linear relationship between x and y. The t-statistic for testing this hypothesis is calculated as:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

where $SE(\hat{\beta}_1)$ is the standard error of the estimated slope. The t-statistic is compared to a critical value from the t-distribution to determine significance.

2.6 Prediction

The regression model can be used to make predictions for the dependent variable given new values of the independent variable. The predicted value $\hat{y}$ for a new observation x_{new} is:

$$\hat{y}_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}$$

3 Conclusion

Linear regression is a fundamental tool in statistical analysis and predictive modeling. By understanding its components and steps, one can effectively apply it to various forecasting and data analysis tasks.