UNIT- IV

Multiple Linear Regression Analysis - II

Prof. Vaishali Poriya

Assumptions of Multiple Linear Regression variables

Multiple linear regression is a statistical method we can use to understand the relationship between multiple predictor variables and a response variable.

Before we perform multiple linear regression, we must first make sure that five assumptions are met:

1. **Linear relationship:** There exists a linear relationship between each predictor variable and the response variable.

2. **No Multicollinearity:** None of the predictor variables are highly correlated with each other.

3. **Independence:** The observations are independent.

4. **Homoscedasticity:** The residuals have constant variance at every point in the linear model.

5. **Multivariate Normality:** The residuals of the model are normally distributed.

NOTE: If one or more of these assumptions are violated, then the results of the multiple linear regression may be unreliable.

Read More at: https://www.statology.org/multiple-linear-regression-assumptions/

Introduction to Multicollinearity

Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other.

Causes of Multicollinearity in Regression Analysis:
1. Correlation Among Predictor Variables:
2. **Overparameterization of the Model**:
3. **Data Collection Issues**:

Detecting Multicollinearity in Regression Analysis
Correlation Matrices:
Variance Inflation Factors (VIFs):

How to mitigate Multicollinearity In Regression Analysis?
Remove Highly Correlated Predictors:
Principal Component Analysis (PCA):

No Multicollinearity

Multiple linear regression assumes that none of the predictor variables are highly correlated with each other.

When one or more predictor variables are highly correlated, the regression model suffers from multicollinearity, which causes the coefficient estimates in the model to become unreliable.

Homoscedasticity

Multiple linear regression assumes that the residuals have constant variance at every point in the linear model. When this is not the case, the residuals are said to suffer from heteroscedasticity.

When heteroscedasticity is present in a regression analysis, the results of the regression model become unreliable.

Specifically, heteroscedasticity increases the variance of the regression coefficient estimates, but the regression model doesn't pick up on this. This makes it much more likely for a regression model to declare that a term in the model is statistically significant, when in fact it is not.

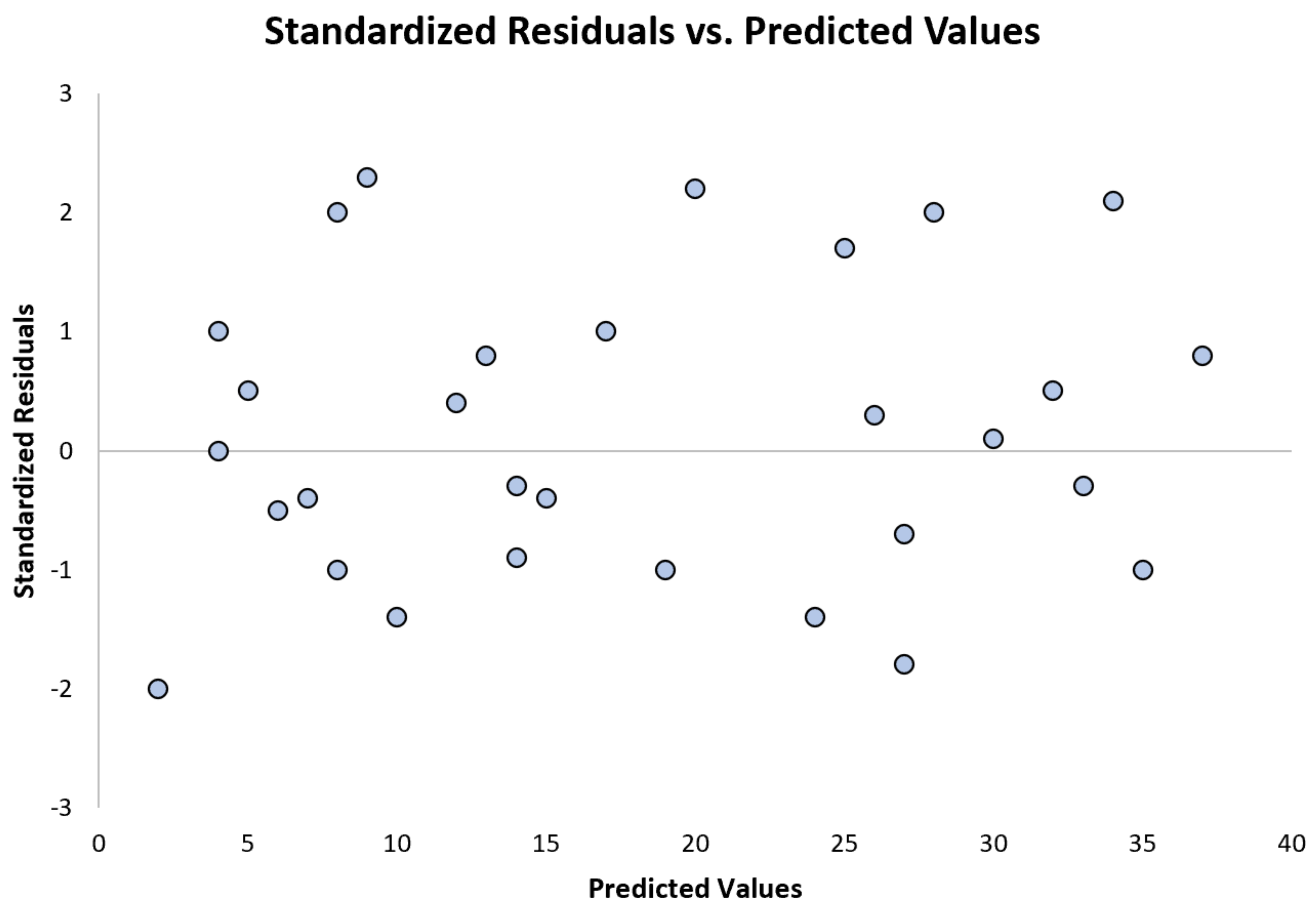
How to Determine if this Assumption is Met

The simplest way to determine if this assumption is met is to create a plot of standardized residuals versus predicted values.

Once you fit a regression model to a dataset, you can then create a scatter plot that shows the predicted values for the response variable on the x-axis and the standardized residuals of the model on the y-axis.

If the points in the scatter plot exhibit a pattern, then heteroscedasticity is present.

The following plot shows an example of a regression model where heteroscedasticity is not a problem:



What to Do if this Assumption is Violated?

1. Transform the response variable. The most common way to deal with heteroscedasticity is to transform the response variable by taking the log, square root, or cube root of all of the values of the response variable. This often causes heteroscedasticity to go away.

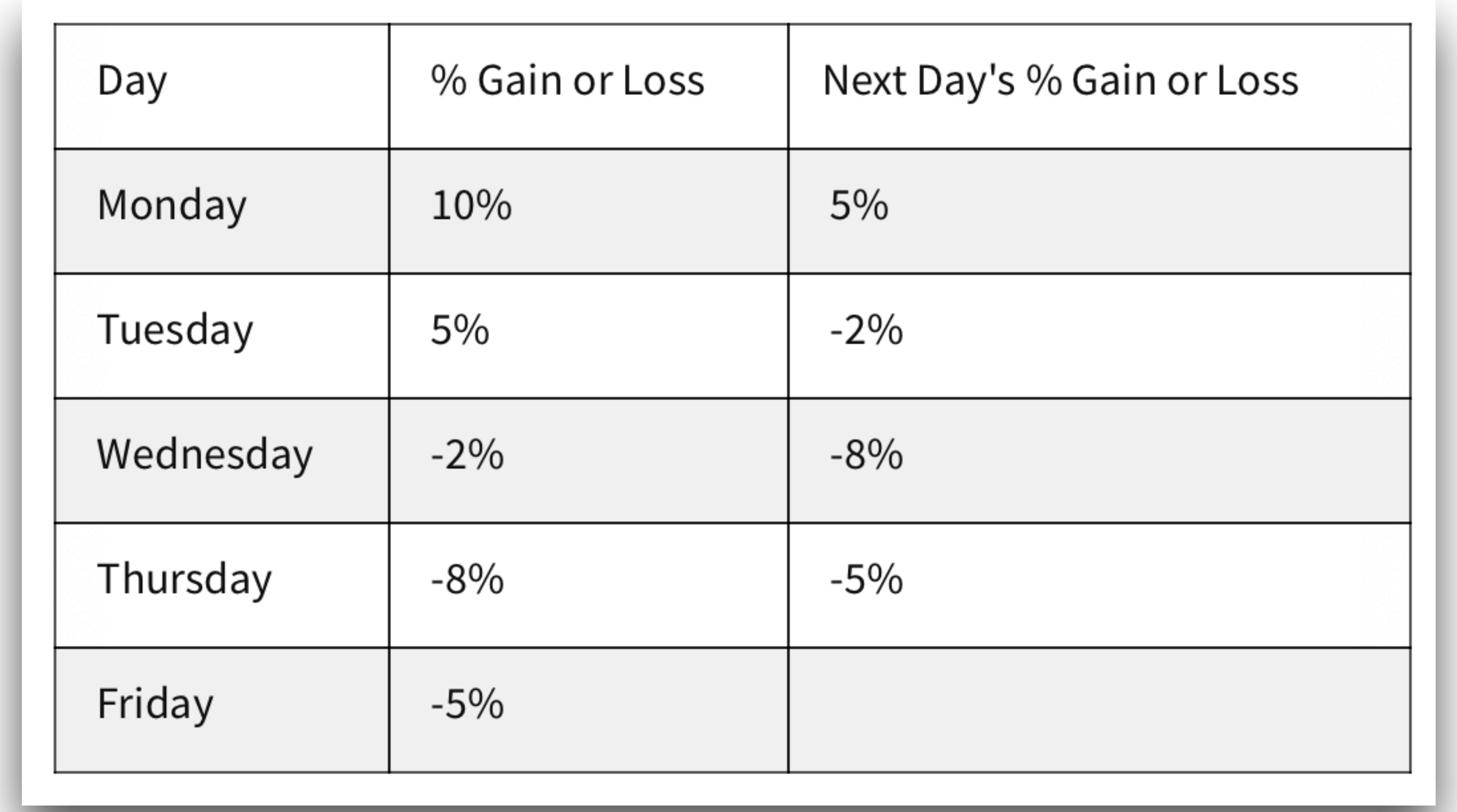
2. Redefine the response variable. One way to redefine the response variable is to use a *rate*, rather than the raw value. For example, instead of using the population size to predict the number of flower shops in a city, we may instead use population size to predict the number of flower shops per capita.

3. Use weighted regression. Another way to fix heteroscedasticity is to use weighted regression, which assigns a weight to each data point based on the variance of its fitted value.

Autocorrelation

Autocorrelation is the degree of similarity of a variable between two successive time intervals.

- Autocorrelation represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals.
- Autocorrelation measures the relationship between a variable's current value and its past values.
- An autocorrelation of +1 represents a perfect positive correlation, while an autocorrelation of -1 represents a perfect negative correlation.
- Technical analysts can use autocorrelation to measure how much influence past prices for a security have on its future price.



Day	% Gain or Loss	Next Day's % Gain or Loss
Monday	10%	5%
Tuesday	5%	-2%
Wednesday	-2%	-8%
Thursday	-8%	-5%
Friday	-5%	

Correlation vs. Autocorrelation

Correlation measures the relationship between two variables, whereas autocorrelation measures the relationship of a variable with lagged values of itself.

What Is the Difference Between Autocorrelation and Multicollinearity?

Autocorrelation is the degree of correlation of a variable's values over time. Multicollinearity occurs when independent variables are correlated and one can be predicted from the other. An example of autocorrelation includes measuring the weather for a city on June 1 and the weather for the same city on June 5. Multicollinearity measures the correlation of two independent variables, such as a person's height and weight.