# Regression Modelling

## Assumptions, Implications and Diagnostics

**Uzair Ahmad**

## Linear Regression Assumptions

For the results of a linear regression model to be valid and reliable, several key assumptions must be met. The key assumptions of linear regression are:

1. **Linearity:** The relationship between predictors and the response variable is linear.
2. **IID:** Residuals are independent of each other and identically distributed.
3. **Homoscedasticity:** Residuals have constant variance.
4. **Normality of Residuals:** Residuals are normally distributed.
5. **No Multicollinearity:** Predictors are not too highly correlated with each other.

Ensuring these assumptions are met helps in deriving reliable and interpretable regression results. If these assumptions are violated, model performance and inference may be compromised, and alternative methods or adjustments might be necessary.

Here's an overview of these assumptions:

### 1. Linearity

The relationship between the dependent variable and the independent variables is linear. This means that changes in the predictor variables are associated with proportional changes in the response variable.

**Implications:**

- The model assumes that the effect of each predictor on the dependent variable is additive and linear.
- To check this assumption, you can use scatterplots of the dependent variable against each predictor or residuals vs. fitted values plots. A linear trend or lack of patterns in these plots supports the linearity assumption.

### 2. IID

**Independence:** Each data point or observation is independent of the others. This means that the occurrence or value of one observation does not affect or is not affected by the occurrence or value of another observation.

**Identically Distributed:** Each data point or observation is drawn from the same probability distribution. This implies that all observations have the same probability distribution with the same parameters.

**Implications:**

- For linear regression models, the IID assumption typically applies to the residuals (errors). The assumption states that residuals should be independently and identically distributed, which is essential for the validity of ordinary least squares (OLS) estimates and hypothesis tests.

### 3. Homoscedasticity

The residuals have constant variance across all levels of the independent variables. This means that the spread or dispersion of the residuals should be similar for all values of the predictors.

**Implications:**

- If residuals display a pattern (e.g., funnel shape) in a residuals vs. fitted values plot, it indicates heteroscedasticity (non-constant variance), which can affect the efficiency of the estimates.
- To address heteroscedasticity, you might need to transform the dependent variable or use weighted least squares regression.

### 4. Normality of Residuals

The residuals (errors) of the model are normally distributed. This assumption is particularly important for constructing confidence intervals and conducting hypothesis tests.

**Implications:**

- Normality is best checked using a QQ plot of the residuals or statistical tests like the Shapiro-Wilk test.
- If residuals are not normally distributed, the estimates of the regression coefficients may still be unbiased, but the standard errors and confidence intervals could be invalid. Transformations of the response variable or robust regression techniques can be considered to address this.

### 5. No Multicollinearity

In multiple linear regression, the independent variables should not be too highly correlated with each other. High correlation between predictors can lead to multicollinearity, which can make the coefficient estimates unstable and inflate their standard errors.

**Implications:**

- Multicollinearity is assessed using variance inflation factors (VIF) or correlation matrices. A high VIF value (typically greater than 10) indicates problematic multicollinearity.
- To address multicollinearity, consider removing or combining correlated predictors, or applying regularization techniques like Ridge or Lasso regression.

# Diagnostic plots

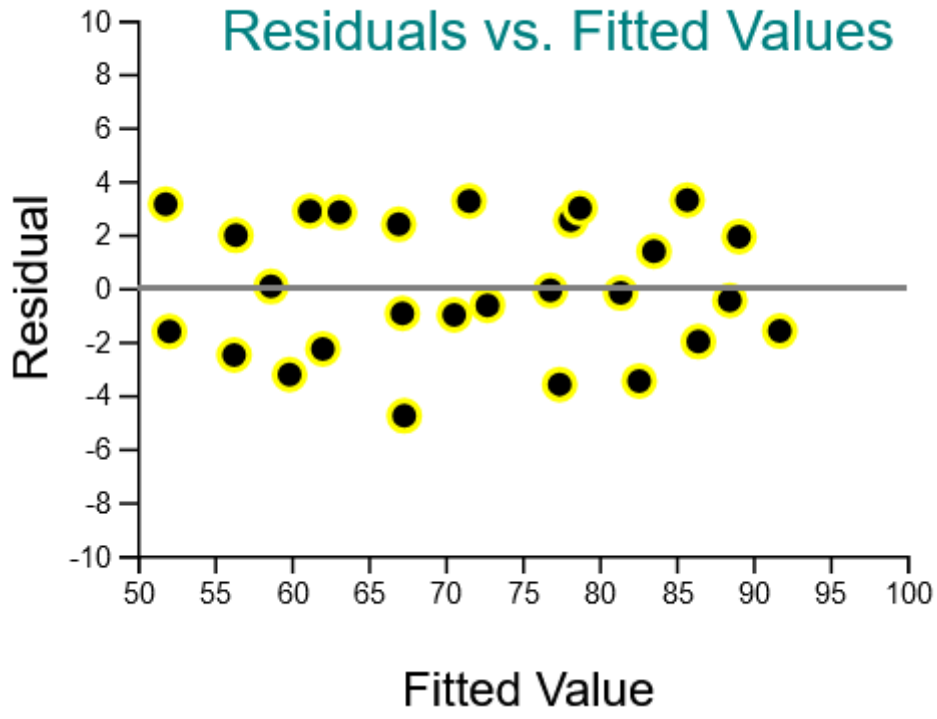
In statistical modeling and regression analysis, it's crucial to evaluate the assumptions of your model to ensure its validity and reliability. Two key diagnostic plots used for this purpose are the **QQ Plot** and the **Residuals vs. Actual Values Plot**. Each provides unique insights into the model's performance and helps identify potential issues.

### Residuals vs. Fitted Values Plot

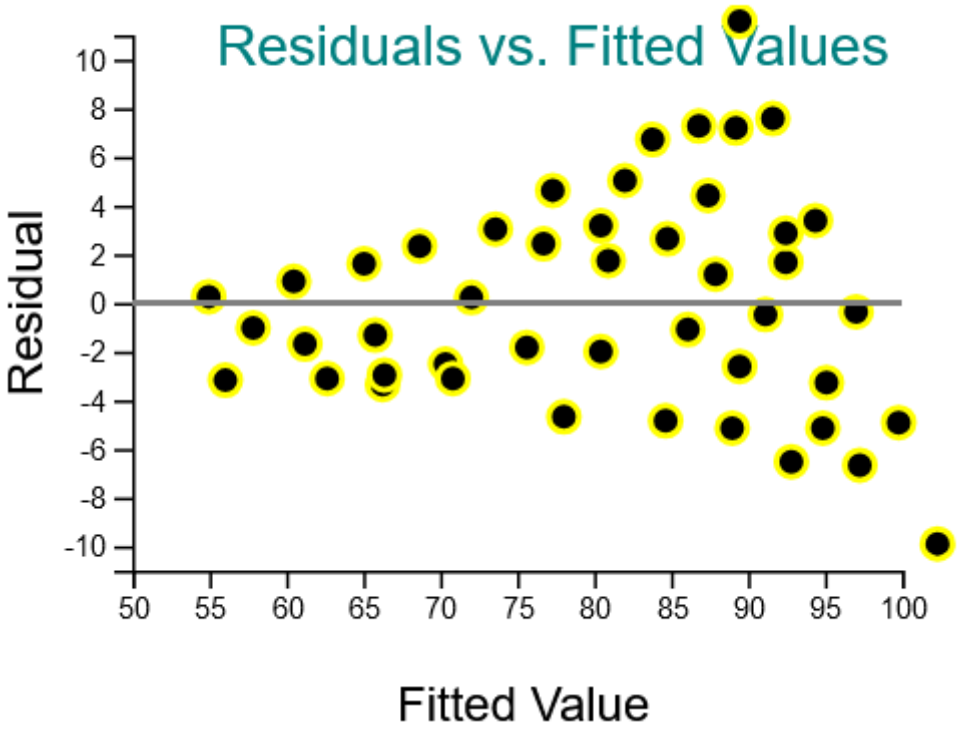
**Purpose:**
The Residuals vs. Fitted Values Plot helps to evaluate whether the residuals exhibit any patterns when plotted against the fitted values. This plot is useful for diagnosing issues such as heteroscedasticity (non-constant variance) and non-linearity in the model.

**How to Use:**

1. **Generate the Plot:** Plot the residuals on the y-axis and the fitted values or actual values on the x-axis. The goal is to examine the spread of residuals across the range of fitted values.
2. **Interpret the Plot:**
    - **Random Scatter:** Ideally, residuals should be randomly scattered around zero without any discernible pattern. This indicates that the model is well-specified and homoscedastic (constant variance).



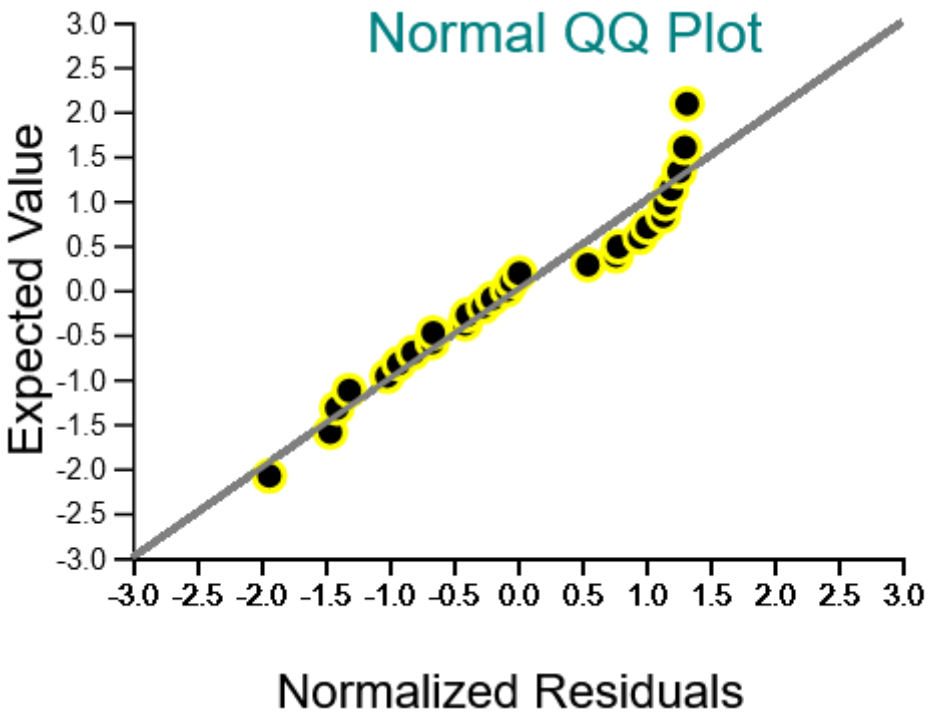

    - **Patterns or Trends:** Systematic patterns, such as a funnel shape (indicating increasing variance) or a curved pattern (indicating non-linearity), suggest issues with the model. These patterns can signal that the model might not be capturing some aspect of the data or that a transformation of variables might be needed.



**Example:**
Suppose you build a regression model to predict student exam scores based on study hours and attendance. After plotting the residuals against the predicted exam scores, you notice a funnel shape (residuals increase with higher predicted values). This indicates heteroscedasticity, suggesting that the variance of the residuals increases with the level of the predicted exam scores. You might need to consider a transformation or use robust regression techniques to address this issue.

## QQ Plot (Quantile-Quantile Plot)

**Purpose:**
The QQ Plot is used to assess whether the residuals (errors) of your model follow a normal distribution. This is an important assumption in many regression models, especially linear regression, where normality of residuals can affect the validity of hypothesis tests and confidence intervals.




**How to Use:**

1. **Generate the QQ Plot:** Plot the quantiles of the residuals against the quantiles of a standard normal distribution. If the residuals are normally distributed, the points should lie approximately along a straight line (usually the 45-degree line).
2. **Interpret the Plot:**
    - **Straight Line Pattern:** If the points closely follow the 45-degree line, it indicates that the residuals are approximately normally distributed.
    - **S-shaped or Curved Pattern:** Deviations from the straight line suggest that the residuals deviate from normality, which could imply model mis-specification or the presence of outliers.

**Example:**
In a linear regression model predicting housing prices, you generate a QQ Plot of the residuals. If the points follow the diagonal line, it confirms that your residuals are normally distributed, validating the assumptions of the linear model.