307102
Descriptive Statistics for Business

Introduction to Linear Regression

2024-2

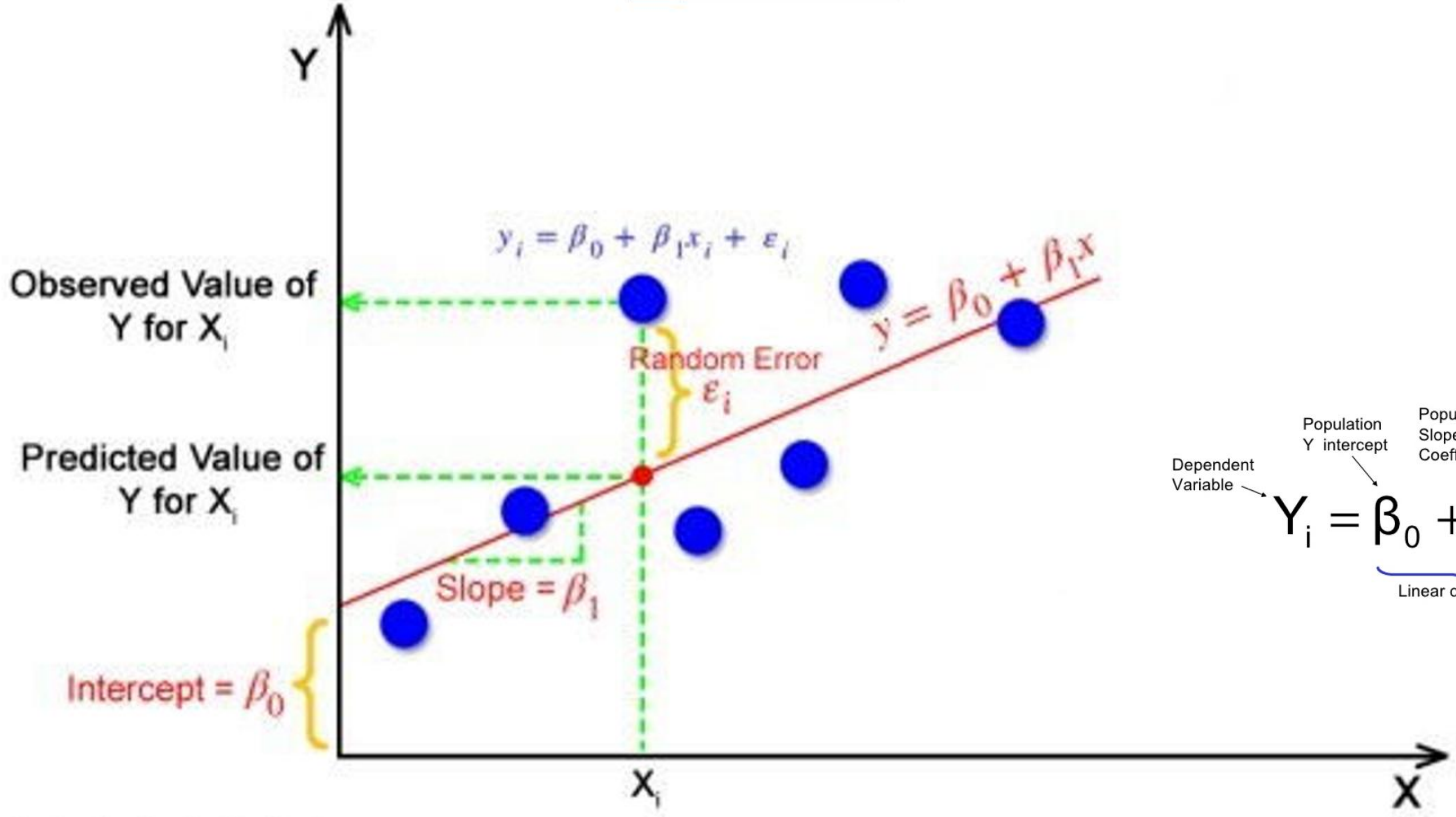
Introduction to Regression Analysis

- Definition: Regression analysis is a powerful statistical method used for predicting a dependent variable based on one or more independent variables.
- Purpose: To understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed

Types of Regression

- Simple Linear Regression: Involves one independent variable and one dependent variable and the relationship between them is modeled by a linear function.
- Multiple Linear Regression (Focus of this presentation): Involves multiple independent variables influencing a single dependent variable.
- Other Types: Logistic regression, polynomial regression, Ridge Regression, Lasso Regression etc., used for more specific types of data and relationships.

The Regression Equation



Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

Finding the Best Fit Line

- The goal of the linear regression algorithm is to get the best values for B0 and B1 to find the best fit line.
- The best fit line is the line that has the least error which means the error between predicted values and actual values should be minimum.
- In regression, the difference between the observed value of the dependent variable(Yi) and the predicted value(predicted) is called the residuals.

- *εi = Y Predicted – Yi*
- *Where Y Predicted = B0 + B1 * Xi*

Mean Squared Error (MSE)

The MSE cost function is given by:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

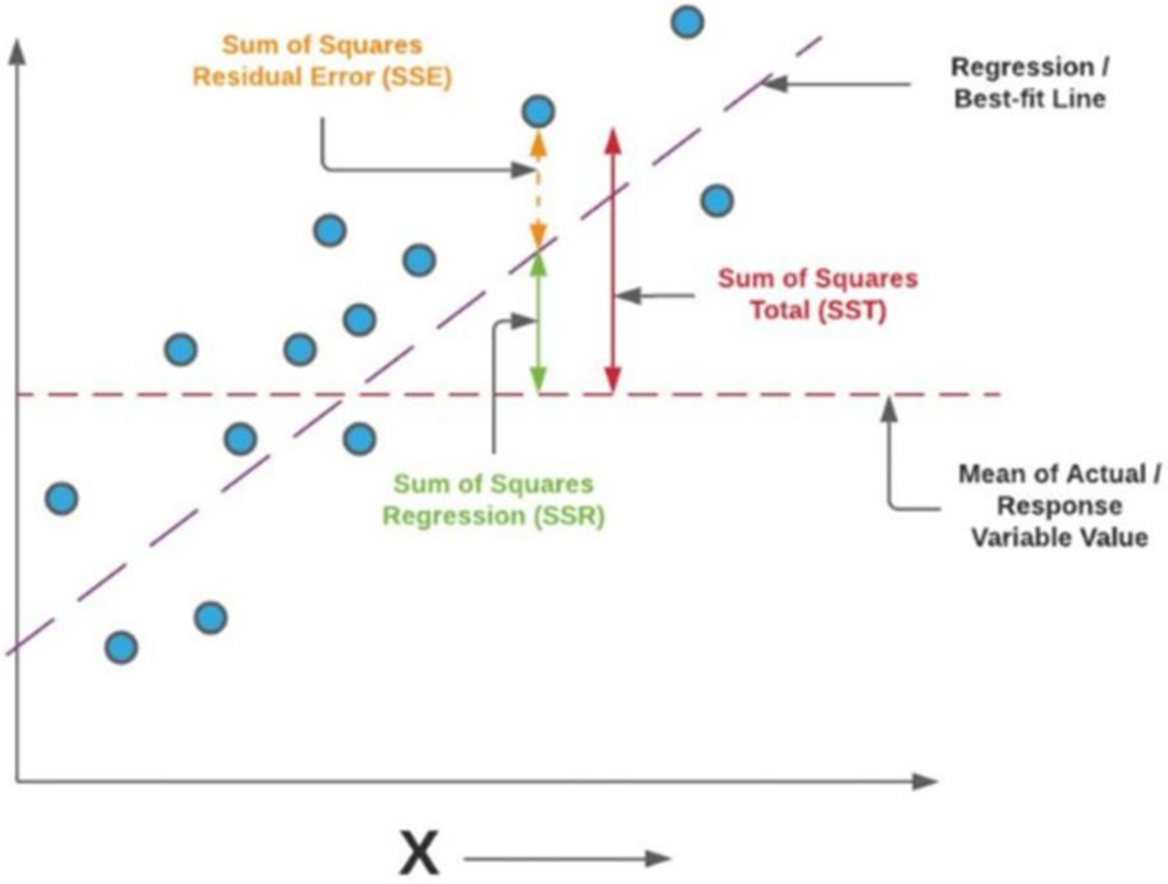
Where:

- $J(\theta)$ is the cost function.
- m is the number of training examples.
- $h_\theta(x^{(i)})$ is the predicted value (hypothesis) for the i-th example.
- $y^{(i)}$ is the actual value for the i-th example.
- $x^{(i)}$ is the feature vector for the i-th example.
- θ represents the parameters (coefficients) of the linear regression model.

- In simple terms, the best fit line is a line that fits the given scatter plot in the best way. Mathematically, the best fit line is obtained by minimizing the Residual Sum of Squares(RSS).

Finding the Best Fit Line

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$



$$a = \frac{[(\sum y)(\sum x^2) - (\sum x)(\sum xy)]}{[n(\sum x^2) - (\sum x)^2]}$$

$$b = \frac{[n(\sum xy) - (\sum x)(\sum y)]}{[n(\sum x^2) - (\sum x^2)]}$$

Mean Squared Error (MSE)

The MSE cost function is given by:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

Where:

- $J(\theta)$ is the cost function.
- m is the number of training examples.
- $h_\theta(x^{(i)})$ is the predicted value (hypothesis) for the i-th example.
- $y^{(i)}$ is the actual value for the i-th example.
- $x^{(i)}$ is the feature vector for the i-th example.
- θ represents the parameters (coefficients) of the linear regression model.

Evaluating the Model

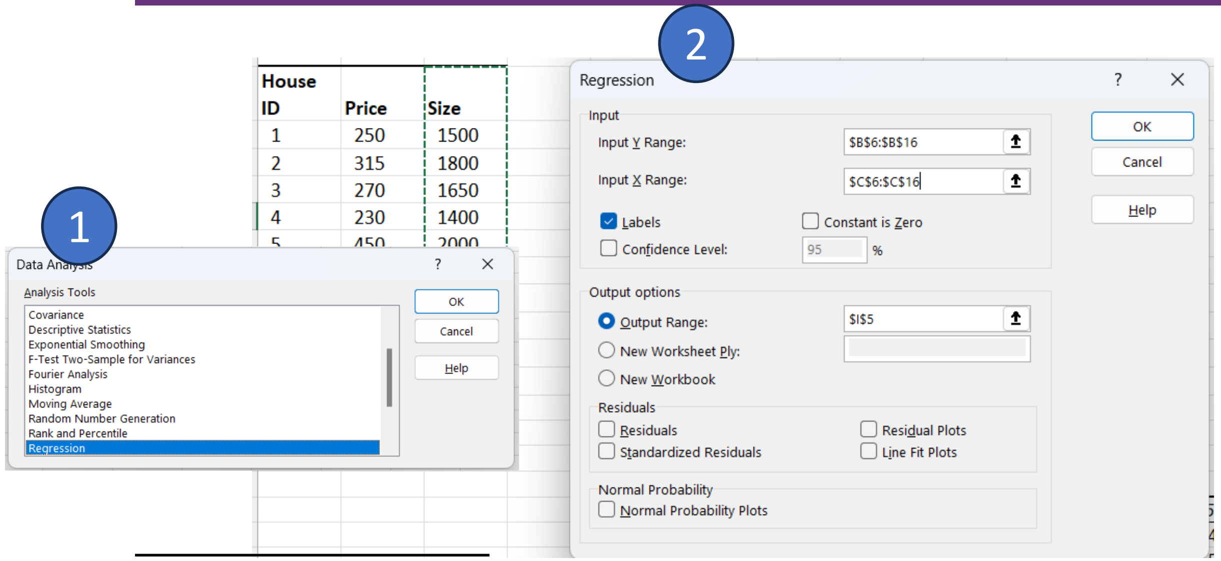
- R-squared (The Goodness of Fit): Proportion of variance in the dependent variable that is predictable from the independent variables.
- Adjusted R-squared: Modified version of R-squared adjusted for the number of predictors.
- F-Statistic: Tests whether at least one predictor variable has a non-zero coefficient.
- t-Tests on Coefficients: Each beta coefficient has an associated t-test to determine if the variable is statistically significant.

Linear Regression Example

- Scenario: A real estate company wants to predict the selling price of houses based on various features.
- Variables:
  - Price: Selling price of the house (in thousands of dollars).
  - Size: Size of the house in square feet.
  - Bedrooms: Number of bedrooms.
  - Age: Age of the house in years.
  - Location: Categorical variable indicating the neighborhood quality (1 = Low, 2 = Medium, 3 = High).

Institute of Analytics
Recognized Programme
University of Petra

Simple Linear Regression



House ID	Price	Size
1	250	1500
2	315	1800
3	270	1650
4	230	1400
5	450	2000
6	190	1300
7	350	1850
8	300	1600
9	280	1500
10	310	1700

3

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.946288
R Square	**0.895461**
Adjusted R	0.882394
Standard E	24.46997
Observatio	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	1	41032.27	41032.27	68.52653	3.41185E-05
Residual	8	4790.234	598.7793		
Total	9	45822.5			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	-217.422	62.32294	-3.48863	0.008215	-361.1388328	-73.7049	-361.139	-73.7049
Size	0.314063	0.037939	8.278075	3.41E-05	0.226574839	0.40155	0.226575	0.40155

Regression Statistics

1. **Multiple R (0.946288)**:
   - This is the **correlation coefficient**, showing the strength and direction of the linear relationship between the independent variable (Size) and the dependent variable (Price).
   - Value close to 1 or -1 indicates a strong linear relationship.
2. **R Square (0.895461)**:
   - The **coefficient of determination**, indicating how much of the variation in the dependent variable (Price) is explained by the independent variable (Size).
   - In this case, ~89.55% of the variance in price is explained by size.
3. **Adjusted R Square (0.882394)**:
   - Adjusted for the number of predictors in the model.
   - Useful when comparing models with different numbers of predictors; penalizes for adding non-significant predictors.
4. **Standard Error (24.46997)**:
   - The **standard deviation of the residuals**. Smaller values indicate better model fit, as residuals are closer to zero.
5. **Observations (10)**:
   - The number of data points used in the regression analysis.

ANOVA (Analysis of Variance):

The ANOVA table tests whether the regression model as a whole is statistically significant.

1. **df (Degrees of Freedom)**:
   - Regression: Number of predictors (1) + Intercept (0) = 1.
   - Residual: Total observations (10) - Number of predictors (1) - 1 = 8.
   - Total: n−1=9n - 1 = 9n−1=9.
2. **SS (Sum of Squares)**:
   - **Regression SS (41032.27)**: Variation in the dependent variable explained by the model.
   - **Residual SS (4790.234)**: Variation not explained by the model (errors).
   - **Total SS (45822.5)**: Total variation in the dependent variable.
3. **MS (Mean Square)**:
   - **Regression MS (41032.27)**: Regression SS divided by its df: 41032.27 ÷ 1
   - **Residual MS (598.7793)**: Residual SS divided by its df: 4790.234 ÷ 8
4. **F (68.52653)**:
   - The **F-statistic**, which tests whether the regression model explains a significant proportion of the variation in the dependent variable.
   - Larger values indicate the model is a good fit.
5. **Significance F (3.41185E-05)**:
   - The **p-value** for the F-test. A small value (e.g., <0.05) indicates the model is statistically significant.

Coefficients Table

1. **Intercept (-217.422)**:
   1. The value of the dependent variable (Price) when the independent variable (Size) is zero.
2. **Size Coefficient (0.31406)**:
   1. The **slope** of the regression line. For every unit increase in size, the price increases by 0.31406 units.
3. **Standard Error (62.32294 for Intercept, 0.037939 for Size)**:
   1. Measures the variability in the coefficient estimates. Smaller values indicate more precise estimates.
   2. The standard error of a coefficient measures the variability in the estimate of that coefficient if we repeatedly sampled data from the population and fitted the regression model each time.
4. **t Stat (-3.48863 for Intercept, 8.278075 for Size)**:
   1. The **t-test statistic**, used to test whether the coefficient is significantly different from zero.
5. **P-value (0.008215 for Intercept, 3.411E-05 for Size)**:
   1. Tests the null hypothesis that the coefficient is zero (no effect).
   2. A small p-value (<0.05) means the coefficient is statistically significant.
   3. The smaller the value, the better the predictor – note value like 9.3499E-5 = 9.3 X 10 ^-5 = 0.00000934
6. **Lower 95% and Upper 95%**:
   1. The **confidence interval** for the coefficient. For example, the coefficient for Size (0.31406) is likely between 0.22657 and 0.40155, with 95% confidence.

Conclusion

- The model is statistically significant (Significance F < 0.05), and **Size** has a significant positive effect on Price (p-value for Size is very small).
- The model explains ~89.55% of the variation in Price ($R^2 = 0.895461$), which is strong.
- The residuals have a standard error of 24.47, indicating the average prediction error.

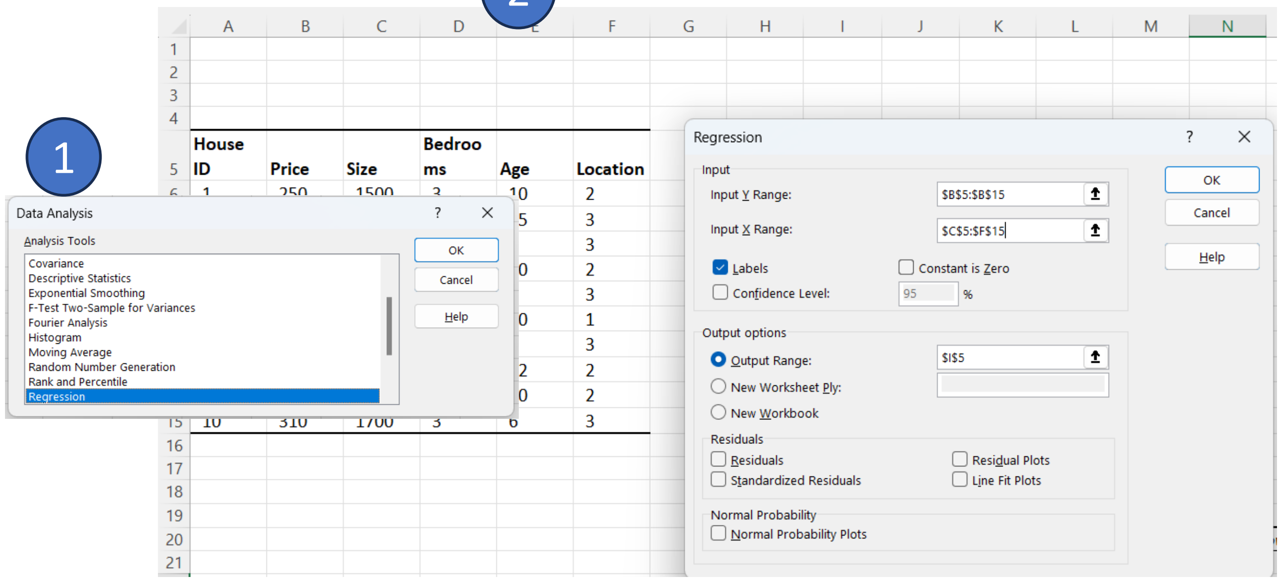
What Does Standard Error Mean in Regression?

- In regression, the **standard error (SE)** reported in the output refers to the **standard deviation of the residuals**, also known as the **residual standard error (RSE)** or simply the "standard error of the regression." It measures how far the observed data points (dependent variable values) are from the predicted values, on average.
- In our example, the standard error value indicates that, on average, the predictions made by your regression model are **off by 24.47 units** (the units of your dependent variable, likely apartment prices).
- It tells you how much variability (or error) remains unexplained by your model.

Multiple Linear Regression

2

1



1	250	1500	3	10	2
2	315	1800	4	15	3
3	270	1650	3	5	3
4	230	1400	2	20	2
5	450	2000	4	2	3
6	190	1300	3	30	1
7	350	1850	4	8	3
8	300	1600	3	12	2
9	280	1500	3	10	2
10	310	1700	3	6	3

Overall Model Accuracy

3

Overall model p-value (significance), small value → good model

Smaller P-Value → Best Predictor

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.990811
R Square	**0.981707**
Adjusted R	0.967073
Standard E	12.94776
Observatic	10

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	4	44984.28	11246.07	67.08281	0.000156
Residual	5	838.2229	167.6446		
Total	9	45822.5			

	Coefficients	*tandard Errc*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	-322.337	73.17067	-4.40527	0.006987	-510.428	-134.246	-510.428	-134.246
Size	0.56681	0.075552	7.502238	0.000665	0.372597	0.761023	0.372597	0.761023
Bedrooms	-36.5822	14.88472	-2.4577	0.057391	-74.8445	1.680218	-74.8445	1.680218
Age	-1.38428	0.964258	-1.4356	0.210597	-3.86299	1.094419	-3.86299	1.094419
Location	-72.3608	15.29873	-4.72986	0.005197	-111.687	-33.0342	-111.687	-33.0342

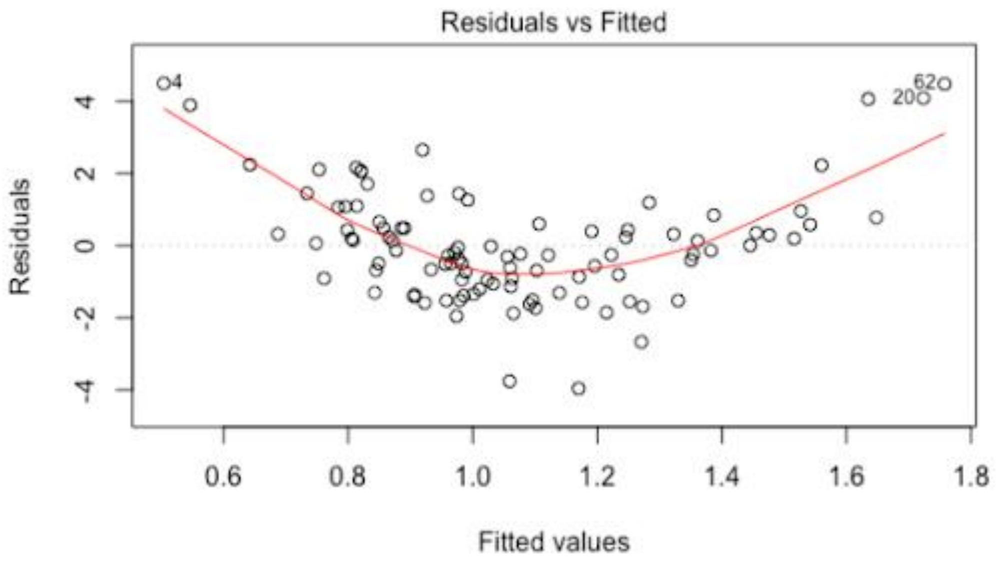
Assumptions of Linear Regression

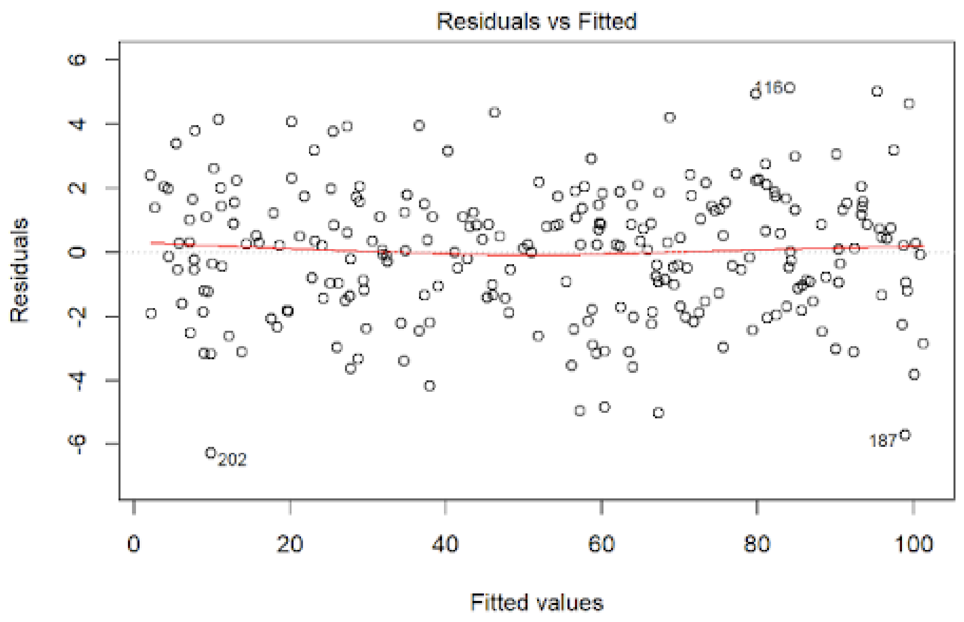
1. Linearity

- The relationship between the independent variables (predictors) and the dependent variable (response) should be linear. This means that the change in the dependent variable is proportional to the change in the independent variable.
- **Example**: If you are predicting house prices (dependent variable) based on the size of the house (independent variable), a linear relationship implies that each additional square foot of the house size increases the price by a constant amount.

Assumptions of Linear Regression

- **How to Test Linearity**: Plot the residuals vs. predicted using a scatter plot, look for a random distribution (no patterns).





To check if the relationship between the independent variable (Size) and the dependent variable (Price) is linear, we look for patterns in the residuals, a random scatter indicates a linear relationship while curves or trends suggest a non-linear relationship.

Assumptions of Linear Regression

2. Independence

- Observations should be independent of each other. This means that the residuals (errors) should not be correlated across observations. This assumption is important for the validity of standard statistical tests.
- Example: In a study measuring the effect of a new drug on blood pressure, each patient's measurement should be independent. If measurements from the same patient at different times are used, they are not independent.
- **How to Test**: Use the Durbin-Watson test to check for autocorrelation in residuals, particularly for time-series data. This test calculates a test statistic (range 0-4). A value around 2 indicates no autocorrelation, values <2 suggest positive autocorrelation, and >2 indicate negative autocorrelation.

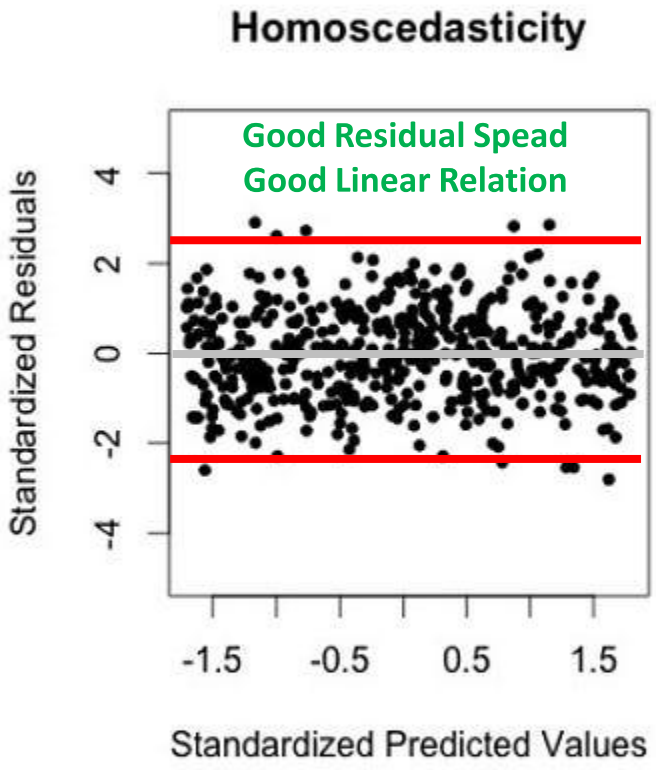
Assumptions of Linear Regression

3. Homoscedasticity

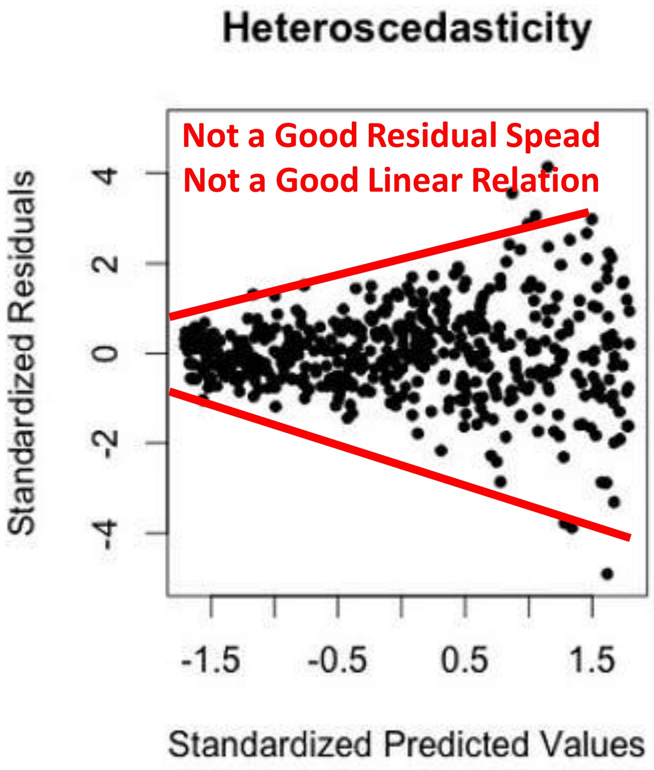
- The variance of the error terms (residuals) should be constant across all levels of the independent variables. If this assumption is violated, it indicates heteroscedasticity, meaning the spread of residuals varies at different levels of the independent variables.
- Example: When plotting residuals against fitted values in a regression analysis of apartment prices against apartment size, the spread of residuals should be roughly constant. If residuals spread out more as apartment size increases, this indicates heteroscedasticity.

Assumptions of Linear Regression

- **How to Test Homoscedasticity** : Plot residuals vs. predicted values. The spread of residuals variance should be consistent (no funnel-shaped patterns).



التوزيع العشوائي للنقاط حول الصفر على طول المدى هو
التوزيع الأفضل الذي يبين أن الموديل الخطي مناسب



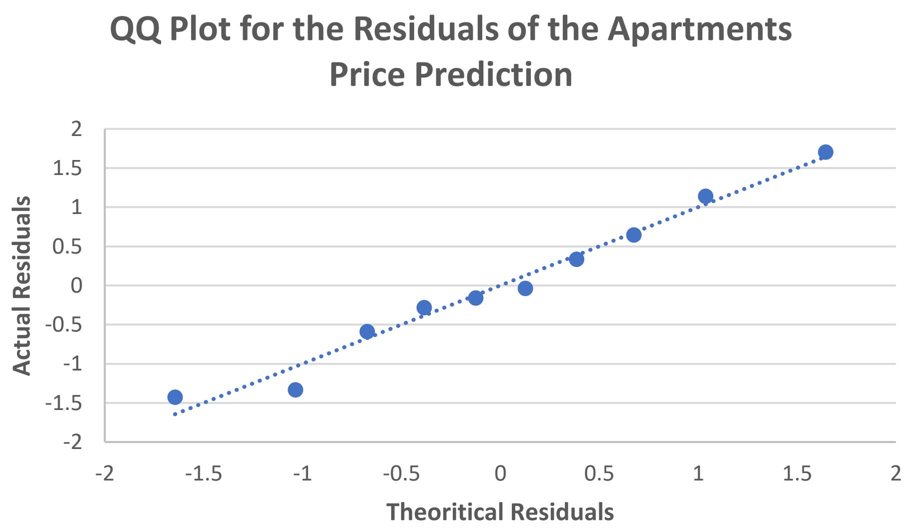
Funnel shape note good
الشكل مثل القمع يبين بأن العلاقة الخطية غير مناسبة

Assumptions of Linear Regression

4. Normal Distribution of Errors

- The error terms should be normally distributed. This assumption is especially important for small sample sizes because it affects the validity of confidence intervals and hypothesis tests.
- Example: After fitting a regression model predicting apartment price based on apartment size, plotting a histogram or a Q-Q plot of the residuals should show that they follow a normal distribution.

Residuals	Standard Residuals	Rank	Percentiles	Theoritical Residuals
-32.890625	-1.425656831	1	0.05	-1.644853627
-30.78125	-1.334225158	2	0.15	-1.036433389
-13.59375	-0.589226339	3	0.25	-0.67448975
-6.484375	-0.281067736	4	0.35	-0.385320466
-3.671875	-0.159158839	5	0.45	-0.125661347
-0.859375	-0.037249941	6	0.55	0.125661347
7.734375	0.335249469	7	0.65	0.385320466
14.921875	0.646794429	8	0.75	0.67448975
26.328125	1.141202737	9	0.85	1.036433389
39.296875	1.703338209	10	0.95	1.644853627



Assumptions of Linear Regression

5. No Multicollinearity

- Independent variables should not be too highly correlated with each other. High correlation between predictors (multicollinearity) can inflate the variances of the coefficient estimates and make the model unstable.
- Example: In a regression model predicting salary based on education level and years of experience, if education level and years of experience are highly correlated, it could cause multicollinearity issues. Checking the Variance Inflation Factor (VIF) can help detect this.
- **How to Test**: Calculate Variance Inflation Factor (VIF). VIF values above 10 (or sometimes 5) indicate high multicollinearity. A VIF >10 (or sometimes >5) indicates problematic multicollinearity.