

307102

Descriptive Statistics for Business

Introduction to Linear Regression

2024-2

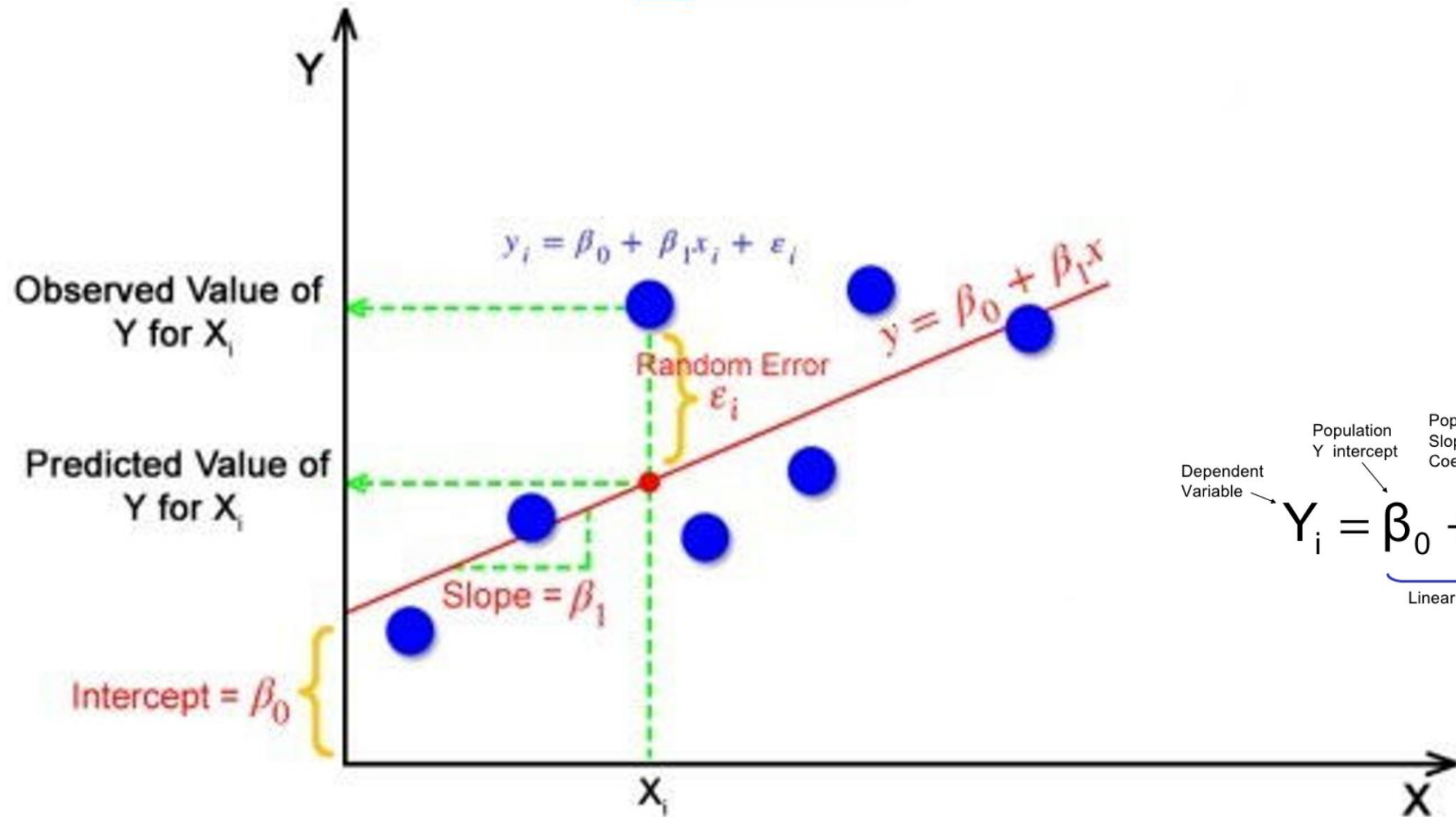
Introduction to Regression Analysis

- 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

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Linear component

Random Error component

Finding the Best Fit Line

- The goal of the linear regression algorithm is to get the best values for B_0 and B_1 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 (Y_i) and the predicted value (predicted) is called the residuals.

- $\epsilon_i = Y_{\text{Predicted}} - Y_i$

- Where $Y_{\text{Predicted}} = B_0 + B_1 * X_i$

- 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).

Mean Squared Error (MSE)

The MSE cost function is given by:

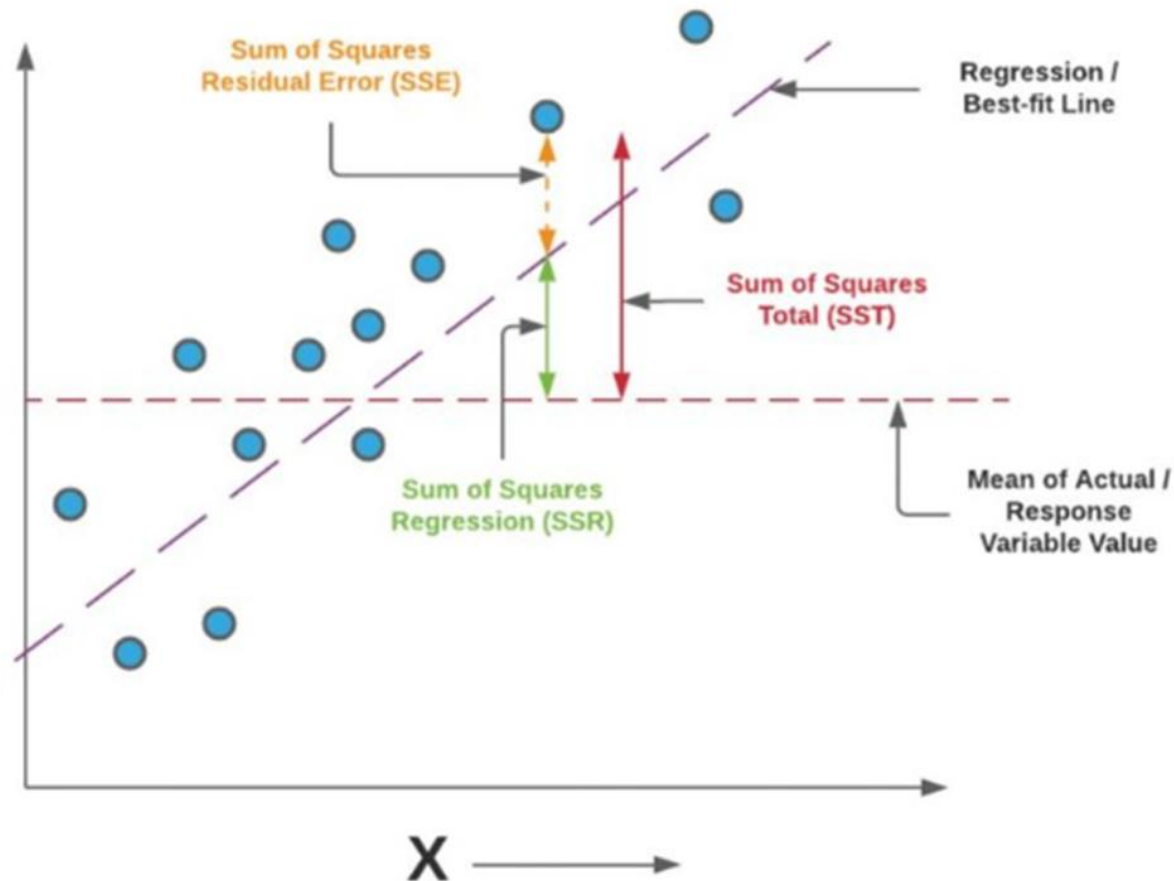
$$J(\theta) = \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

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.

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).

Simple Linear Regression

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

3

SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.946288								
R Square	0.895461								
Adjusted R	0.882394								
Standard E	24.46997								
Observations	10								
ANOVA									
	df	SS	MS	F	Significance F				
Regression	1	41032.27	41032.27	68.52653	3.41185E-05				
Residual	8	4790.234	598.7793						
Total	9	45822.5							
		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 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: $n-1=9$ - 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: $41032.27 \div 1$
- **Residual MS (598.7793)**: Residual SS divided by its df: $4790.234 \div 8$

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 $9.3499\text{E-}5 = 9.3 \times 10^{-5} = 0.00000934$

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

1

2

Data Analysis

Analysis Tools

- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression

OK

Cancel

Help

Regression

Input Y Range: \$B\$5:\$B\$15

Input X Range: \$C\$5:\$F\$15

☒ Labels

☐ Constant is Zero

Confidence Level: 95 %

Output options

☒ Output Range: \$I\$5

☐ New Worksheet Ply:

☐ New Workbook

Residuals

☐ Residuals

☐ Standardized Residuals

☐ Residual Plots

☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK

Cancel

Help

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

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.990811
R Square	0.981707
Adjusted R Square	0.967073
Standard Error	12.94776
Observations	10

ANOVA

	df	SS	MS	F	Significance F
Regression	4	44984.28	11246.07	67.08281	0.000156
Residual	5	838.2229	167.6446		
Total	9	45822.5			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 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

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

Smaller P-Value → Best Predictor

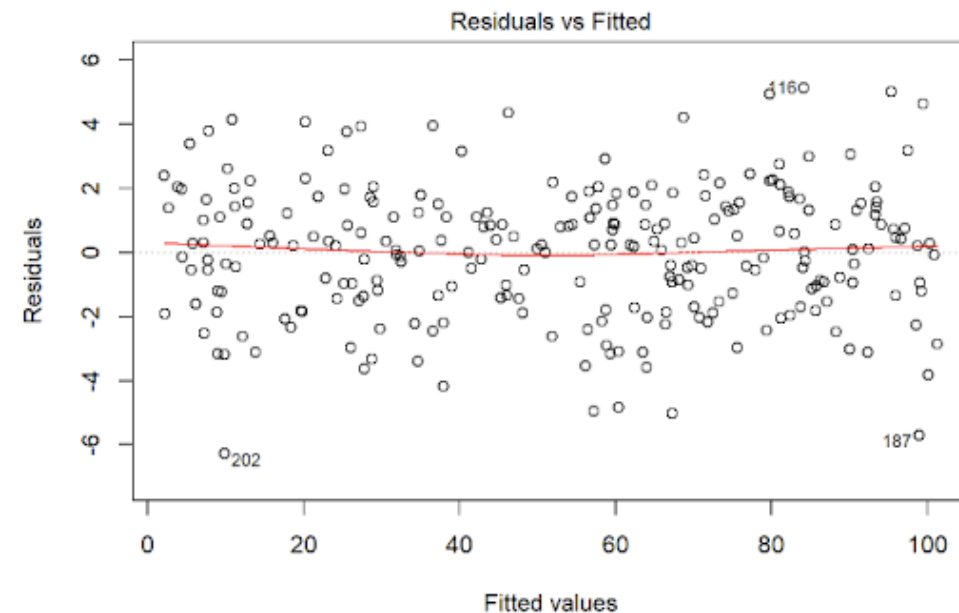
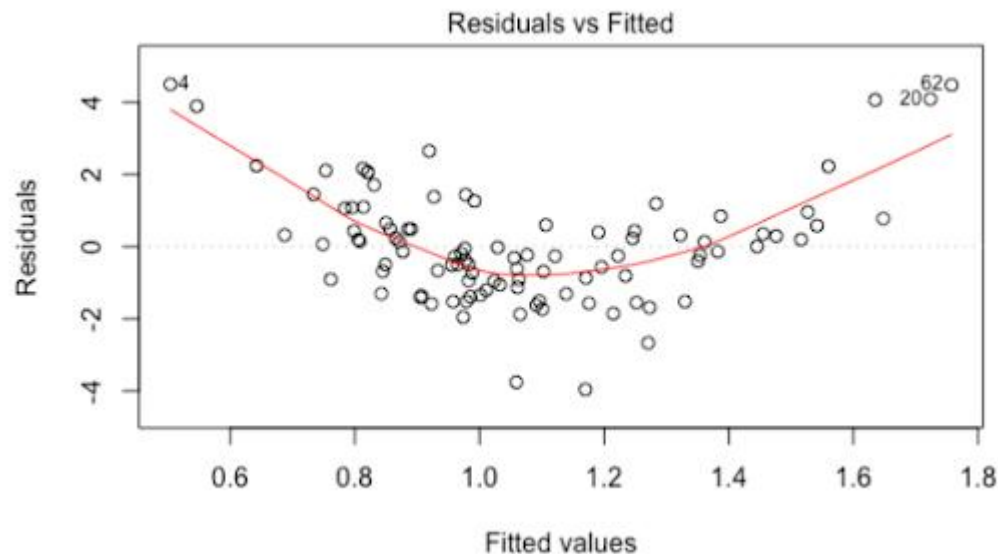
Assumptions of Linear Regression

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.

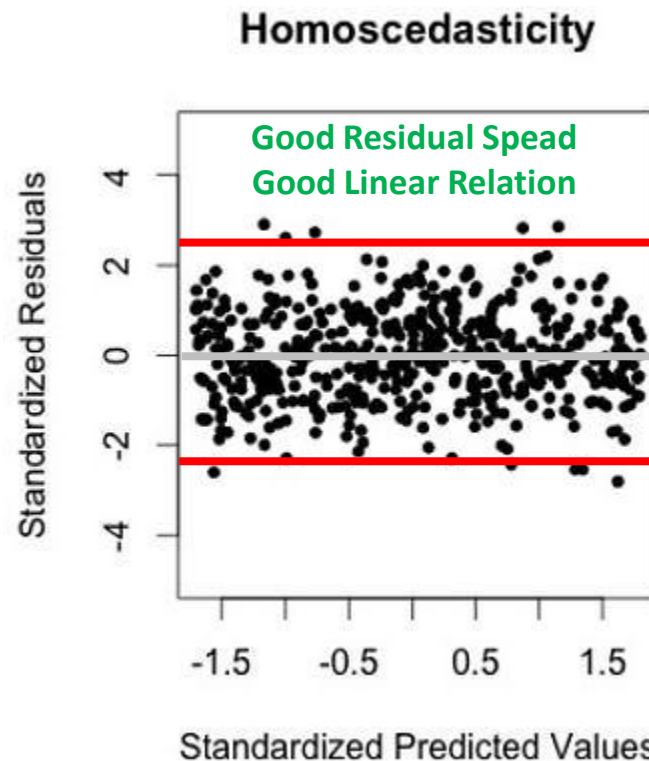
Assumptions of Linear Regression

3. Homoscedasticity

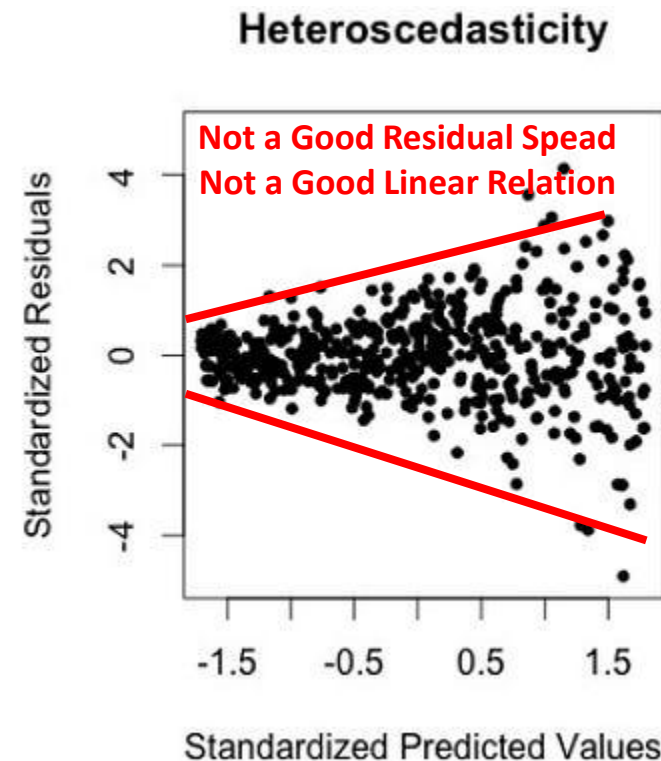
- The variance of the error terms (residuals) should be constant across all levels of the independent variables. If this assumption is violated, it indicates heteroscedasticity, meaning the spread of residuals varies at different levels of the independent variables.
- Example: When plotting residuals 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 النموذج الخطي مناسب



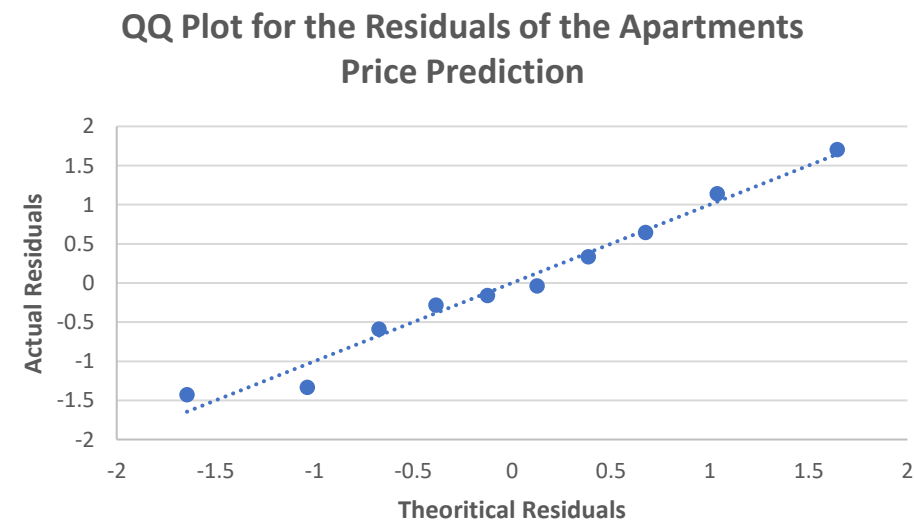
Funnel shape note good
الشكل مثل القمع يبين بأن العلاقة الخطية غير مناسبة

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	Theoretical 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.