Linear Regression Errors Explained

Linear Regression: Theory, Problems, and Solutions

1. Theory of Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable Y and one or more independent variables X. It is commonly used for predictive modeling. The equation of a simple linear regression model is:

$$Y = mX + c + \epsilon$$

Where:

- Y is the dependent variable (response)
- X is the independent variable (predictor)
- m (or β_1) is the slope of the line
- c (or β_0) is the intercept
- ϵ is the error term (unexplained variance)

For multiple regression, the equation extends to:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

where multiple independent variables ($X_1, X_2, ...X_n$) influence Y.

Error Metrics in Linear Regression

To measure how well a linear regression model fits the data, various error metrics are used:

1. **Mean Absolute Error (MAE):**

   $$MAE = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \hat{Y}_i|$$

   Measures the average absolute difference between actual and predicted values.

2. **Mean Squared Error (MSE):**

   $$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

   Measures the average squared differences between actual and predicted values.

3. **Root Mean Squared Error (RMSE):**

$$RMSE = \sqrt{MSE}$$

   Provides a more interpretable measure by taking the square root of MSE.

4. **R-Squared (R^2):**

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

   Measures the proportion of variance in Y explained by X. A value close to 1 indicates a better fit.

Problem 1: Predicting House Prices Based on Size

Given Data:

House Size (sq. ft)	Price (in $1000s)
1400	245
1600	312
1700	279
1875	308
1100	199

Solution:

We need to fit a linear regression model:

$$Price = m \times \text{Size} + c$$

Step 1: Calculate Mean Values

$$\bar{X} = \frac{1400 + 1600 + 1700 + 1875 + 1100}{5} = 1535$$

$$\bar{Y} = \frac{245 + 312 + 279 + 308 + 199}{5} = 268.6$$

Step 2: Calculate Slope m and Intercept c

$$m = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$m = \frac{(1400 - 1535)(245 - 268.6) + (1600 - 1535)(312 - 268.6) + (1700 - 1535)(279 - 268.6) +}{(1400 - 1535)^2 + (1600 - 1535)^2 + (1700 - 1535)^2 + (1875}$$

$$m = \frac{(-135)(-23.6) + (65)(43.4) + (165)(10.4) + (340)(39.4) + (-435)(-69.6)}{(-135)^2 + (65)^2 + (165)^2 + (340)^2 + (-435)^2}$$

$$m = \frac{3186 + 2821 + 1716 + 13396 + 30216}{18225 + 4225 + 27225 + 115600 + 189225}$$

$$m = \frac{51335}{354500} = 0.145$$

$$c = \bar{Y} - m \times \bar{X}$$

$$c = 268.6 - (0.145 \times 1535)$$

$$c = 268.6 - 222.6 = 46$$

Thus, the regression equation is:

$$Price = 0.145 \times \text{Size} + 46$$

Step 3: Predicting and Evaluating Errors

Using this model, we predict prices and compute errors.

House Size	Actual Price	Predicted Price	Error (Actual - Predicted)	Squared Error
1400	245	0.145(1400)+46 = 248.3	-3.3	10.89
1600	312	0.145(1600)+46 = 277.2	34.8	1211.04
1700	279	0.145(1700)+46 = 291.7	-12.7	161.29
1875	308	0.145(1875)+46 = 318.2	-10.2	104.04
1100	199	0.145(1100)+46 = 205.9	-6.9	47.61

MAE:

$$MAE = \frac{|-3.3| + |34.8| + |-12.7| + |-10.2| + |-6.9|}{5}$$

$$MAE = \frac{67.9}{5} = 13.58$$

MSE:

$$MSE = \frac{10.89 + 1211.04 + 161.29 + 104.04 + 47.61}{5}$$

$$MSE = \frac{1534.87}{5} = 306.97$$

RMSE:

$$RMSE = \sqrt{306.97} = 17.52$$

R-Squared Calculation:

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

$$R^2 = 1 - \frac{1534.87}{5602.8} = 0.726$$

Thus, the model explains **72.6%** of the variance in house prices.

Problem 2: Predicting Student Scores Based on Study Hours

Given Data:

Study Hours (X)	Exam Score (Y)
1.5	55
3.0	60
4.5	68
5.0	72
6.0	80

We will fit a linear regression model:

$$Y = mX + c$$

Step 1: Calculate Mean Values

$$\bar{X} = \frac{1.5 + 3.0 + 4.5 + 5.0 + 6.0}{5} = 4.0$$

$$\bar{Y} = \frac{55 + 60 + 68 + 72 + 80}{5} = 67.0$$

Step 2: Calculate Slope m and Intercept c

The formula for slope:

$$m = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$m = \frac{(1.5 - 4.0)(55 - 67) + (3.0 - 4.0)(60 - 67) + (4.5 - 4.0)(68 - 67) + (5.0 - 4.0)(72 - 67) +}{(1.5 - 4.0)^2 + (3.0 - 4.0)^2 + (4.5 - 4.0)^2 + (5.0 - 4.0)^2 + (6.0 - 4.0)^2}$$

$$m = \frac{(-2.5)(-12) + (-1.0)(-7) + (0.5)(1) + (1.0)(5) + (2.0)(13)}{(-2.5)^2 + (-1.0)^2 + (0.5)^2 + (1.0)^2 + (2.0)^2}$$

$$m = \frac{30 + 7 + 0.5 + 5 + 26}{6.25 + 1.0 + 0.25 + 1.0 + 4.0}$$

$$m = \frac{68.5}{12.5} = 5.48$$

Now, calculate the intercept:

$$c = \bar{Y} - m \times \bar{X}$$

$$c = 67 - (5.48 \times 4)$$

$$c = 67 - 21.92 = 45.08$$

Thus, the regression equation is:

$$\text{Score} = 5.48 \times \text{Study Hours} + 45.08$$

Step 3: Predicting Scores and Errors

Study Hours	Actual Score	Predicted Score $5.48X + 45.08$	Error (Actual - Predicted)	Squared Error
1.5	55	$5.48(1.5) + 45.08 = 53.4$	$55 - 53.4 = 1.6$	2.56
3.0	60	$5.48(3.0) + 45.08 = 61.5$	$60 - 61.5 = -1.5$	2.25
4.5	68	$5.48(4.5) + 45.08 = 69.6$	$68 - 69.6 = -1.6$	2.56
5.0	72	$5.48(5.0) + 45.08 = 72.5$	$72 - 72.5 = -0.5$	0.25
6.0	80	$5.48(6.0) + 45.08 = 77.9$	$80 - 77.9 = 2.1$	4.41

Step 4: Compute Error Metrics

Mean Absolute Error (MAE):

$$MAE = \frac{|1.6| + |1.5| + |1.6| + |0.5| + |2.1|}{5}$$

$$MAE = \frac{7.3}{5} = 1.46$$

Mean Squared Error (MSE):

$$MSE = \frac{2.56 + 2.25 + 2.56 + 0.25 + 4.41}{5}$$

$$MSE = \frac{12.03}{5} = 2.41$$

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{MSE} = \sqrt{2.41} = 1.55$$

R-Squared (R^2) Calculation:

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Total variation:

$$\sum(Y_i - \bar{Y})^2 = (55-67)^2 + (60-67)^2 + (68-67)^2 + (72-67)^2 + (80-67)^2$$

$$= (-12)^2 + (-7)^2 + (1)^2 + (5)^2 + (13)^2$$

$$= 144 + 49 + 1 + 25 + 169 = 388$$

Residual variation:

$$\sum(Y_i - \hat{Y}_i)^2 = 12.03$$

$$R^2 = 1 - \frac{12.03}{388} = 0.969$$

Thus, the model explains **96.9%** of the variance in exam scores.

Final Summary of Results:

Problem 1: Predicting House Prices

- **Regression Equation:** $Y = 0.145X + 46$
- **MAE:** 13.58
- **MSE:** 306.97
- **RMSE:** 17.52
- **R^2:** 72.6%

Problem 2: Predicting Exam Scores

- **Regression Equation:** $Y = 5.48X + 45.08$
- **MAE:** 1.46
- **MSE:** 2.41
- **RMSE:** 1.55
- **R^2:** 96.9%

Conclusion

- The **house price model** has an R^2 of **72.6%**, meaning it explains 72.6% of price variations.
- The **exam score model** has an R^2 of **96.9%**, indicating a **very strong correlation**.
- Both models perform well, but the exam score model fits better due to lower error values and higher R^2.