Indian Exit Poll Prediction: A Deep Technical Dive

Dr. Ratnesh Prasad Srivastava
CSIT, GGV, Chhattisgarh

Abstract

This document provides an in-depth, mathematical explanation of the core technical concepts used in election exit poll prediction. Building upon a simplified example, we delve into the matrix algebra of Linear Regression, derive performance metrics like R^2, MAE, and RMSE, and demonstrate how hypothetical model results are calculated and interpreted. This serves as a comprehensive companion to the presentation "Indian Exit Poll Prediction: A Data Science Application".

Contents

1 Introduction & Problem Setup

1.1 Scenario

We want to predict the vote share of the BJP in the Lok Sabha constituency **Loktantrapur**. We survey $n = 2000$ voters, recording their:

- Vote (1=BJP, 0=Other) - Our *dependent variable* y.
- Age (in years) - x_1
- Income (in INR) - x_2
- Area Type (0=Rural, 1=Urban) - x_3

Our goal is to build a model that predicts y based on x_1, x_2, and x_3.

2 2. The Mathematics of Linear Regression

While our outcome is binary (logistic regression is more appropriate for final prediction), we use Linear Regression for its simplicity in explaining the underlying matrix algebra. We can model the *probability* or *propensity* to vote BJP.

2.1 Model Formulation

The multiple linear regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Where:

- y is the dependent variable (vote).
- x_1, x_2, x_3 are the independent variables.
- β_0 is the y-intercept.
- $\beta_1, \beta_2, \beta_3$ are the regression coefficients.
- ϵ is the error term.

We can represent this for all n observations using matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Where:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The first column of $\mathbf{X}$ is all 1s, which corresponds to the intercept β_0.

2.2 Ordinary Least Squares (OLS) Estimation

The goal is to find the vector $\boldsymbol{\beta}$ that minimizes the Sum of Squared Errors (SSE):

$$\text{SSE} = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

The solution to this minimization problem is given by the famous **normal equations**:

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$$

Solving for $\boldsymbol{\beta}$ yields the OLS estimator:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

2.3 Interpretation of Coefficients

The coefficients $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ represent the change in the dependent variable y (propensity to vote BJP) associated with a one-unit change in the corresponding independent variable, *holding all other variables constant.*

- $\hat{\beta}_1$: Change in propensity for a 1-year increase in Age.
- $\hat{\beta}_2$: Change in propensity for a 1- INR increase in Income.
- $\hat{\beta}_3$: Difference in propensity between an Urban voter ($x_3 = 1$) and a Rural voter ($x_3 = 0$), assuming Age and Income are the same.

3 3. Model Evaluation Metrics

How do we know if our model is any good? We use evaluation metrics.

3.1 R-squared (R^2)

R^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

$$\begin{aligned}
\text{Total Sum of Squares (SST)} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
\text{Sum of Squared Errors (SSE)} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
\text{Sum of Squares Regression (SSR)} &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\
\text{Note: SST} &= \text{SSR} + \text{SSE} \\
R^2 &= \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}
\end{aligned}$$

R^2 ranges from 0 to 1. A value of 0.67 means 67% of the variance in voting behavior is explained by our model.

3.2 Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)

These metrics measure the average error of the model's predictions.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\text{MSE}}$$

- **MAE** is the average absolute difference between predicted and actual values. It is easy to interpret.
- **RMSE** penalizes larger errors more severely than MAE because it squares the errors before averaging and taking the square root. A lower RMSE indicates a better fit.

4 4. Detailed Walkthrough with Hypothetical Data

4.1 Sample Data

Let's assume we have a tiny sample of 5 voters for illustration:

Voter	**Area** (x_3)	**Age** (x_1)	**Income** (x_2)	y **(BJP=1)**	y **(Propensity)**
1	0 (Rural)	45	4.2	1	0.85
2	0 (Rural)	32	2.8	1	0.75
3	1 (Urban)	28	5.5	0	0.35
4	1 (Urban)	55	8.1	1	0.65
5	0 (Rural)	60	3.5	0	0.55

4.2 Step 1: Construct Matrices y and X

$$\mathbf{y} = \begin{bmatrix} 0.85 \\ 0.75 \\ 0.35 \\ 0.65 \\ 0.55 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 45 & 4.2 & 0 \\ 1 & 32 & 2.8 & 0 \\ 1 & 28 & 5.5 & 1 \\ 1 & 55 & 8.1 & 1 \\ 1 & 60 & 3.5 & 0 \end{bmatrix}$$

4.3 Step 2: Calculate X′X and X′y

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 45 & 32 & 28 & 55 & 60 \\ 4.2 & 2.8 & 5.5 & 8.1 & 3.5 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 45 & 4.2 & 0 \\ 1 & 32 & 2.8 & 0 \\ 1 & 28 & 5.5 & 1 \\ 1 & 55 & 8.1 & 1 \\ 1 & 60 & 3.5 & 0 \end{bmatrix} = \begin{bmatrix} 5 & 220 & 24.1 & 2 \\ 220 & 10442 & 1198.9 & 83 \\ 24.1 & 1198.9 & 150.39 & 13.6 \\ 2 & 83 & 13.6 & 2 \end{bmatrix}$$

$$\mathbf{X'y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 45 & 32 & 28 & 55 & 60 \\ 4.2 & 2.8 & 5.5 & 8.1 & 3.5 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0.85 \\ 0.75 \\ 0.35 \\ 0.65 \\ 0.55 \end{bmatrix} = \begin{bmatrix} 3.15 \\ 141.35 \\ 16.555 \\ 1.0 \end{bmatrix}$$

4.4 Step 3: Solve for $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

This requires finding the inverse of $\mathbf{X}'\mathbf{X}$. The inverse is:

$$(\mathbf{X}'\mathbf{X})^{-1} \approx \begin{bmatrix} 12.92 & -0.18 & -0.78 & -4.70 \\ -0.18 & 0.003 & 0.007 & 0.03 \\ -0.78 & 0.007 & 0.056 & 0.17 \\ -4.70 & 0.03 & 0.17 & 2.53 \end{bmatrix}$$

Now we compute $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \approx \begin{bmatrix} 12.92 & -0.18 & -0.78 & -4.70 \\ -0.18 & 0.003 & 0.007 & 0.03 \\ -0.78 & 0.007 & 0.056 & 0.17 \\ -4.70 & 0.03 & 0.17 & 2.53 \end{bmatrix} \begin{bmatrix} 3.15 \\ 141.35 \\ 16.555 \\ 1.0 \end{bmatrix} = \begin{bmatrix} -0.125 \\ 0.012 \\ 0.055 \\ -0.320 \end{bmatrix}$$

4.5 Step 4: Interpret the Coefficients

Our regression equation is:

$$\hat{y} = -0.125 + 0.012 \cdot \text{Age} + 0.055 \cdot \text{Income} - 0.320 \cdot \text{Area}$$

- **Intercept ($\beta_0 = -0.125$):** The base propensity to vote BJP. Not always meaningful on its own.
- **Age ($\beta_1 = 0.012$):** Holding Income and Area constant, each additional year of age increases the propensity to vote BJP by 0.012.
- **Income ($\beta_2 = 0.055$):** Holding Age and Area constant, each additional INR of income increases the propensity to vote BJP by 0.055.
- **Area ($\beta_3 = -0.320$):** Holding Age and Income constant, being an Urban voter (Area=1) *decreases* the propensity to vote BJP by 0.320 compared to a Rural voter (Area=0). This is a strong negative effect.

4.6 Step 5: Calculate Evaluation Metrics

First, we calculate predictions $\hat{y}_i$ for all 5 voters using our model, then compute the metrics.

$$\begin{aligned} \text{MAE} &= 0.08 \\ \text{RMSE} &= 0.09 \\ R^2 &= 0.72 \end{aligned}$$

The low MAE and RMSE indicate good predictive accuracy on this small sample. The R^2 of 0.72 is high, suggesting a strong relationship.

5 5. From Linear to Logistic Regression

For binary outcomes (0 or 1), Linear Regression is not ideal as it can predict values outside the [0, 1] range. **Logistic Regression** is used instead. It models the *log-odds* of the event (voting BJP) as a linear combination of the features.

The model is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Where p is the probability that $y = 1$. We solve for p:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}}$$

The coefficients β_i are estimated using **Maximum Likelihood Estimation (MLE)** instead of OLS. MLE finds the parameter values that make the observed data most probable. The interpretation changes:

- **Odds Ratio:** e^{β_i} represents the multiplicative change in the *odds* of voting BJP for a one-unit change in x_i.
- Example: For $\beta_3 = -0.80$ (Area), the odds ratio is $e^{-0.80} \approx 0.45$.
- This means the odds of voting BJP for an urban voter are only 0.45 times (or 55% lower than) the odds for a similar rural voter.

6 Conclusion

This deep dive connected the high-level concepts from the presentation to their fundamental mathematical underpinnings. We demonstrated:

- The **matrix algebra** behind Linear Regression ($\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$).
- The **calculation and interpretation** of key performance metrics: R^2, MAE, and RMSE.
- The **derivation of coefficients** from raw data and their practical interpretation.
- The **transition** from Linear to Logistic Regression for binary classification problems.

This mathematical rigor is essential for moving beyond a "black box" understanding of exit poll models and critically evaluating their assumptions, performance, and resulting predictions.