Error Based Regression

HI 743

Ryan Gallagher

Department of Health Informatics and Administration
Zilber College of Public Health
University of Wisconsin - Milwaukee

February 20, 2025

Overview

1. Simple Linear Regression

1.1 Simple Linear Model
1.2 Statistical Notation
1.3 Estimating Coefficients
1.4 Assessing Model Accuracy

2. Multiple Linear Regression

2.1 Multiple Regression Model
2.2 Regression Coefficients
2.3 Collinearity
2.4 Assessing Model Fit
2.5 Feature Selection

Introduction to Simple Linear Regression

Simple Linear Regression is a method used to model the relationship between a **single predictor** variable X and a **quantitative response variable** Y.

Simple Linear Model:

$$Y \approx \beta_0 + \beta_1 X \tag{1}$$

Where the *coefficients*

- β_0 is the intercept (value of Y when $X = 0$).
- β_1 is the slope (change in Y for a one-unit increase in X)

We assume that there is approximately a linear relationship between X and Y. This is an approximation because there exists some error in our prediction.

Simple Linear Regression

Example - X may represent TV Advertising and Y may represent Sales. We can regress Sales onto Ads by fitting the model:

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{Ads}$$

Interpretation: *For every one-unit increase in* `Ads`*, we get a* β_1 *unit increase in* `sales`*, plus some* β_0*.*

Simple Linear Regression

We, however, don't know the true values of β_0 and β_1. Thus, we must estimate these coefficients.

The Prediction Model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2}$$

Where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$.

Note: In machine learning, we use training data to estimate the coefficients, then test how well the model estimates Y in the test data.

Statistical Notation

Notation in our Data:

- x_i – The i-th observed value of the **predictor variable** (independent variable).
- y_i – The i-th observed value of the **response variable** (dependent variable).

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

Notation in the *True* model:

- β_0 – The **intercept**, representing the predicted value of Y when $X = 0$.
- β_1 – The **slope**, representing the change in Y for a one-unit increase in X.

Statistical Notation

Notation in our *Estimated* model:

- $\hat{\beta}_1$ – The **estimated slope** obtained from the data using least squares.
- $\hat{\beta}_0$ – The **estimated intercept** obtained from the data.
- $\hat{y}_i$ – The **predicted value** of y_i, given by:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{3}$$

- e_i – The **residual** (error) for observation i, calculated as:

$$e_i = y_i - \hat{y}_i \tag{4}$$

Estimating the Regression Coefficients

We estimate the coefficients β_0 and β_1 using the **least squares** method.

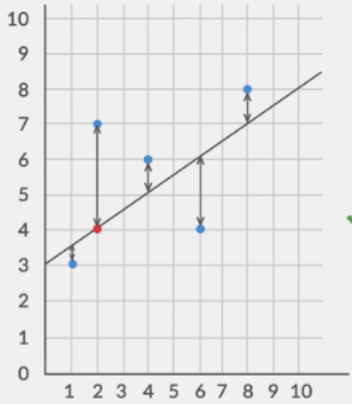
Residual Sum of Squares (RSS):

$$RSS = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{5}$$

$$= \sum_{i=1}^{n} e_i^2 \tag{6}$$

The **Least Squares** approach chooses $\hat{\beta}_0$ & $\hat{\beta}_1$ that minimizes the RSS.

Estimating the Regression Coefficients



$$Y = \beta_0 + \beta_1 X$$

Intercept Slope

$$e_i = y_i - y_{pred}$$

Ordinary Least Squares Method:

$$e_1^2 + e_2^2 + \_\_\_ + e_n^2 = \text{RSS (Residual Sum Of Squares)}$$

$$RSS = (Y_1-\beta_0-\beta_1 X_1)^2 + (Y_2-\beta_0-\beta_1 X_2)^2 + \_\_\_\_ + (Y_n-\beta_0-\beta_1 X_n)^2$$

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Assessing the Accuracy of the Coefficient Estimates

Our estimates have inherent uncertainty due to sample variability.

The Population Regression Model:

(The best linear approximation of the true relationship between X and Y.)

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{7}$$

Where ϵ is a random error term with mean zero. This is a catch-all term for what we miss with this simple model - that the relationship is *probably* not linear.

The best linear model minimizes ϵ

Residual Standard Error (RSE)

RSE measures how much the actual Y values deviate from the regression line. It is similar to the standard deviation but applied to regression errors (residuals).

Residual Standard Error (RSE):

$$RSE = \sqrt{\frac{1}{n-2}\sum(y_i - \hat{y}_i)^2} \tag{8}$$

Interpretation:

- A smaller RSE means the model fits the data well.
- If RSE $= 3$, on average, the predicted values ($\hat{y}$) differ from actual values (y) by about 3 units.

R^2 (R-Squared): Goodness of Fit

R^2 measures how well the model explains variation in Y. R^2 is the proportion of variation in Y that is explained by X.

R^2 (R-Squared):

$$R^2 = 1 - \frac{RSS}{TSS} \tag{9}$$

where:

- RSS = Residual Sum of Squares $\sum(y_i - \hat{y}_i)^2$ (unexplained variation).
- TSS = Total Sum of Squares $\sum(y_i - \bar{y})^2$ (total variation in Y).

Interpretation:

- R^2 ranges from 0 to 1.
- $R^2 = 0.8 \rightarrow 80\%$ of variation in Y is explained by X.
- Higher R^2 means a better model, but beware of overfitting!

Introduction to Multiple Linear Regression

Multiple Linear Regression extends simple linear regression to multiple predictors. It helps us understand how multiple variables impact the response variable.

Multiple Linear Regression Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{10}$$

where:

- Y is the response variable.
- $X_1, X_2, ..., X_p$ are predictor variables.
- $\beta_0, \beta_1, ..., \beta_p$ are unknown coefficients.
- ϵ is the error term.

Example: Predicting Sales

$$\text{Sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{Radio} + \beta_3 \times \text{Newspaper} + \epsilon \tag{11}$$

Estimating the Regression Coefficients

Similar to Simple Linear Regression...

- Use Least Squares to estimate $\beta_0, \beta_1, ..., \beta_p$.
- Choose coefficients that minimize the Residual Sum of Squares (RSS).

RSS Formula:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}))^2 \quad (12)$$

How are β_j values found?

- β_j refer to all variables in your model - requires matrix algebra to solve RSS.
- R and other statistical tools compute these estimates automatically.

Understanding Regression Coefficients

Interpretation of β_j:

- β_j represents the change in Y for a one-unit increase in X_j, **holding all other predictors constant**.
- Example: If $\beta_2 = 0.189$ is our radio ad coefficient, then a \$1,000 increase in radio advertising increases sales by **189 units** (given all other coefficients remain constant).

Caution: Correlation Between Predictors

- Coefficients can change dramatically if predictors are correlated.
- Example: TV and radio spending might be correlated, affecting their individual estimates.

Collinearity in Multiple Regression

Collinearity:

- When predictor variables are **highly correlated**, the regression model becomes unstable.
- Leads to unreliable coefficient estimates and high standard errors.

Detecting Collinearity:

- **Variance Inflation Factor (VIF)** measures multicollinearity:

$$VIF(X_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}} \tag{13}$$

- $VIF > 5$ suggests collinearity issues.

Managing Collinearity

- Remove or combine correlated predictors.
- Use other regression techniques (PCA, Ridge Regression)

Model Fit: Adjusted R^2

Adjusted R^2 improves on R^2 by considering the number of predictors. Unlike R^2, it penalizes unnecessary predictors to prevent overfitting.

Adjusted R^2**:**

$$R^2_{\text{adj}} = 1 - \left(\frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)} \right) \tag{14}$$

Interpretation:

- Higher Adjusted $R^2 \rightarrow$ better fit while controlling for complexity.
- Adding unnecessary predictors **decreases** Adjusted R^2.
- Useful for comparing models with **different numbers of predictors**.

Model Fit: Mean Squared Error (MSE)

MSE measures the average squared difference between actual and predicted values. Lower MSE means the model has **better predictive accuracy**. This is a standard metric when comparing *most* ML models.

Formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{15}$$

Interpretation:

- **Smaller MSE** $\rightarrow$ Model predictions are closer to true values.
- It quantifies how well the model predicts unseen data.

Feature Selection in Regression

Why Feature Selection?

- Not all predictors contribute significantly to the model.
- Removing irrelevant features improves interpretability and reduces overfitting.
- Can enhance prediction accuracy and model efficiency.

Types of Features:

- **Predictive Features** – Directly useful for estimating Y.
- **Interacting Features** – Not useful alone but important with others.
- **Redundant Features** – Strongly correlated with another predictor.
- **Irrelevant Features** – No useful information for predicting Y.

Goal: Select a minimal subset that maintains model performance.

Feature Selection Methods

1. **Filter Methods (Pre-Processing)**
   - Rank features using statistical metrics (e.g., correlation, mutual information).
   - Remove low-ranking features before modeling.

2. **Wrapper Methods (Model-Based)**
   - Iteratively select the best subset of features by training models.
   - Common approaches:
     - **Forward Selection** – Start with none, add features stepwise.
     - **Backward Selection** – Start with all, remove stepwise.
     - **Stepwise Selection** – Combines forward and backward approaches.

3. **Regularization (Shrinkage)**
   - Penalizes large coefficients to reduce complexity.
   - **LASSO (L1)** forces some coefficients to zero (feature selection).
   - **Ridge Regression (L2)** shrinks all coefficients but keeps them nonzero.