Regression for Statistics and Machine Learning

Pranav Viswanath

MATH 3215

Contents

1 Introduction

1.1 What is Regression?

Regression analysis is a fundamental statistical tool that allows us to model and explore relationships between variables. At its core, regression helps us quantify the impact of one or more independent variables on a dependent variable. This technique finds its applications in various fields, providing a means to make predictions, uncover patterns, and understand complex phenomena.

Regression originated in the late 19th century, evolving into a versatile method used in science, business, and public policy. For instance, consider predicting home prices based on features like size, number of bedrooms, and location. Regression enables us to discern the underlying connections between these variables and make informed predictions.

Key Terminology:

- **Dependent Variable:** The variable we aim to predict or explain (e.g., home prices).
- **Independent Variable:** The variable used to predict or explain the dependent variable (e.g., size, number of bedrooms).
- **Linear Regression:** Modeling the relationship between variables with a straight line.
- **Residuals:** Differences between observed and predicted values.

1.2 Historical Context of Regression Analysis

The roots of regression analysis can be traced back to the work of Sir Francis Galton in the 19th century. Galton's pioneering research on heredity and variation laid the foundation for regression. Later, Karl Pearson and Udny Yule further developed the methodology.

In the early 20th century, the concept of least squares estimation, a key component of regression analysis, was formalized. This period marked the establishment of regression analysis as a statistical technique with widespread applications.

Today, regression analysis has evolved significantly, with extensions like multiple regression, polynomial regression, and machine learning-based approaches, making it a cornerstone of data analysis.

1.3 Motivating Example: Predicting Home Prices

To illustrate the power of regression, let's consider a concrete example: predicting home prices. Imagine we have a dataset that includes information about various houses, such as their size, number of bedrooms, and location. Our goal is to build a model that can accurately predict the selling price of a house based on these features.

The regression model, in this case, would look something like this:

$$\text{Price} = \beta_0 + \beta_1 \times \text{Size} + \beta_2 \times \text{Bedrooms} + \beta_3 \times \text{Location} + \varepsilon$$

Here,

- β_0 is the intercept term, representing the base price of a house.
- $\beta_1, \beta_2, \beta_3$ are the coefficients that quantify the impact of each feature on the house price.
- ε is the error term, representing the difference between the predicted and actual prices.

This example showcases how regression allows us to leverage data to make predictions in a real-world context.

2 Types of Regression

2.1 Simple Linear Regression

In simple linear regression, we model the relationship between a dependent variable y and a single independent variable x using the equation:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Here,

- y is the dependent variable we want to predict.
- x is the independent variable.
- β_0 is the y-intercept, representing the value of y when x is 0.
- β_1 is the slope, indicating how much y changes for a unit change in x.
- ε is the error term, accounting for unexplained variability in y.

Let's delve into a step-by-step example to solidify our understanding.

2.1.1 Step-by-Step Example: Fitting a Simple Linear Model

Consider the following dataset: $(1, 3)$, $(2, 5)$, $(3, 8)$, $(4, 10)$. We want to fit a simple linear regression model to this data.

The simple linear regression model has the form:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Our task is to find the values of β_0 and β_1 that minimize the sum of squared differences between the observed and predicted values. This process is often done using the method of ordinary least squares.

Step 1: Calculate Means

$$\bar{x} = \frac{1 + 2 + 3 + 4}{4} = 2.5$$

$$\bar{y} = \frac{3 + 5 + 8 + 10}{4} = 6.5$$

Step 2: Calculate Slope (β_1)

$$\beta_1 = \frac{\sum_{i=1}^{4}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{4}(x_i - \bar{x})^2}$$

$$\beta_1 = \frac{(1 - 2.5)(3 - 6.5) + (2 - 2.5)(5 - 6.5) + (3 - 2.5)(8 - 6.5) + (4 - 2.5)(10 - 6.5)}{(1 - 2.5)^2 + (2 - 2.5)^2 + (3 - 2.5)^2 + (4 - 2.5)^2}$$

After calculating, we find $\beta_1 = 1.5$.

Step 3: Calculate Intercept (β_0)

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_0 = 6.5 - (1.5 \times 2.5) = 1$$

Therefore, the fitted simple linear regression model is:

$$y = 1.5x + 1$$

This equation represents the best-fit line that minimizes the sum of squared differences between the observed and predicted values.

Step 4: R-squared Value The R-squared value measures the goodness of fit of the model. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{4}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{4}(y_i - \bar{y})^2}$$

Where $\hat{y}_i$ is the predicted value for each observation.
Let's calculate this for our example:

$$\hat{y}_1 = 1.5 \times 1 + 1 = 2.5$$
$$\hat{y}_2 = 1.5 \times 2 + 1 = 4$$
$$\hat{y}_3 = 1.5 \times 3 + 1 = 5.5$$
$$\hat{y}_4 = 1.5 \times 4 + 1 = 7$$

Now, plug these values into the R-squared formula:

$$R^2 = 1 - \frac{(3 - 2.5)^2 + (5 - 4)^2 + (8 - 5.5)^2 + (10 - 7)^2}{(3 - 6.5)^2 + (5 - 6.5)^2 + (8 - 6.5)^2 + (10 - 6.5)^2}$$

After calculation, we find the R-squared value to assess the model's goodness of fit.

Conclusion In this example, we learned how to fit a simple linear regression model to a dataset. We calculated the slope (β_1), intercept (β_0), and the R-squared value, providing insights into the strength of the linear relationship between the variables.

2.2 Multiple Linear Regression

While simple linear regression deals with one independent variable, multiple linear regression extends this concept to multiple predictors. The equation is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

Here,

- y is the dependent variable.
- $x_1, x_2, \ldots, x_k$ are the independent variables.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients.
- ε is the error term.

Let's work through an example to understand multiple linear regression in practice.

2.2.1 Step-by-Step Example: Multiple Linear Regression

Consider a dataset with house features: size, number of bedrooms, and location. We want to predict the sale price of a house using these features. The multiple linear regression model is:

$$\text{Price} = \beta_0 + \beta_1 \times \text{Size} + \beta_2 \times \text{Bedrooms} + \beta_3 \times \text{Location} + \varepsilon$$

Our task is to find the values of $\beta_0, \beta_1, \beta_2, \beta_3$ that minimize the sum of squared differences between the observed and predicted prices.

Step 1: Formulate the Model The multiple linear regression model is given by:

$$\text{Price} = \beta_0 + \beta_1 \times \text{Size} + \beta_2 \times \text{Bedrooms} + \beta_3 \times \text{Location} + \varepsilon$$

Here, β_0 is the base price, β_1 represents the impact of size on price, β_2 represents the impact of the number of bedrooms, and β_3 represents the impact of location.

Step 2: Calculate Coefficients To find the coefficients, we use the method of least squares. The formulas for $\beta_0, \beta_1, \beta_2, \beta_3$ involve means, variances, and covariances.

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}$$

$$\beta_2 = \frac{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}$$

$$\beta_3 = \frac{\sum_{i=1}^{n}(x_{3i} - \bar{x}_3)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{3i} - \bar{x}_3)^2}$$

$$\beta_0 = \bar{y} - \beta_1\bar{x}_1 - \beta_2\bar{x}_2 - \beta_3\bar{x}_3$$

Here, x_{1i}, x_{2i}, x_{3i} are the sizes, number of bedrooms, and locations for each house, respectively.

Step 3: Interpret Coefficients Once we have the coefficients, we can interpret them in the context of the problem. For example, if β_1 is positive, it means that an increase in size is associated with an increase in price. If β_2 is negative, it implies that more bedrooms might lead to a lower price.

Step 4: Assess Model Performance We can evaluate the performance of the model using metrics like R-squared. This helps us understand how well our model explains the variability in house prices based on the provided features.

Conclusion In this example, we delved into multiple linear regression, extending our understanding from simple linear regression. We formulated the model, calculated coefficients, interpreted their meaning, and discussed how to assess the model's performance.

2.3 Polynomial Regression

In polynomial regression, the goal is to capture non-linear relationships in the data by introducing polynomial terms. The model equation is given by:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \varepsilon$$

Here,

- y is the dependent variable.
- x is the independent variable.
- $\beta_0, \beta_1, \ldots, \beta_k$ are the coefficients.
- k is the degree of the polynomial.
- ε is the error term.

2.3.1 Example

Consider modeling the relationship between age (x) and salary (y). We can fit polynomials of varying degrees and compare their R-squared values to determine the best fit for the data.

2.3.2 Assessing Performance

The performance of polynomial regression can be assessed using metrics like R-squared, which measures the goodness of fit. Higher R-squared values indicate a better fit to the data.

2.3.3 Conclusion

Polynomial regression is a flexible tool for capturing non-linear patterns in data. However, choosing the appropriate degree of the polynomial is crucial to avoid overfitting or underfitting.

2.4 Support Vector Regression

Support Vector Regression (SVR) uses kernel functions to map data into a higher-dimensional space. The SVR equation is:

$$y = w^T \phi(x) + b$$

Here,

- y is the dependent variable.
- w is the weight vector.
- $\phi(x)$ is the kernel-transformed feature vector.
- b is the bias term.

2.4.1 Example

Suppose we have data on housing prices (y) and features like square footage and location (x). SVR can be applied to predict housing prices by capturing non-linear relationships.

2.4.2 Assessing Performance

Evaluate SVR performance using metrics like Mean Squared Error (MSE) or R-squared. SVR is effective in capturing complex patterns in the data.

2.4.3 Conclusion

SVR is a powerful regression technique for handling non-linear relationships. The choice of kernel function and tuning parameters is crucial for optimal performance.

2.5 Decision Tree Regression

Decision Tree Regression involves recursively partitioning the feature space. The model predicts the dependent variable based on the decision tree structure.

2.5.1 Model Equation

The decision tree equation is a set of rules leading to a predicted value for each observation.

2.5.2 Example

Consider predicting a student's GPA (y) based on study hours and attendance (x). The decision tree branches based on conditions like study hours ¿ 10 and attendance ¿ 90

2.5.3 Assessing Performance

Evaluate decision tree performance using metrics like Mean Absolute Error (MAE) or Mean Squared Error (MSE). Decision trees provide transparency into feature importance.

2.5.4 Conclusion

Decision tree regression is interpretable and effective for capturing complex relationships. However, they are prone to overfitting, and techniques like pruning are used to mitigate this.

2.6 Random Forest Regression

Random Forest Regression is an ensemble method that combines predictions from multiple decision trees.

2.6.1 Model Equation

The random forest aggregates predictions from individual decision trees to obtain a more robust prediction.

2.6.2 Example

Predicting stock prices (y) based on various financial indicators (x) using a random forest model.

2.6.3 Assessing Performance

Random Forest performance is assessed using metrics like Mean Squared Error or Out-of-Bag Error. It helps in mitigating overfitting.

2.6.4 Conclusion

Random Forest Regression is a powerful ensemble technique that improves predictive accuracy and handles high-dimensional data effectively.

3 Philosophy of Regression

Regression analysis has connections to philosophical debates on causality, induction, and the nature of knowledge. On the surface regression is a mathematical technique, but its history touches on deep questions about modeling relationships.

The term "regression" was introduced by Francis Galton to describe biological phenomena where offspring tend to have average, central characteristics compared to the population. This opened up discussions on whether regression reveals causal mechanisms or is merely descriptive shorthand. Karl Pearson critiqued regression as tautologous, while Ronald Fisher argued it could determine causal impacts.

A core philosophical issue is the problem of induction - how can broad generalizations be reliably deduced from limited data observations? As David Hume discussed, inductive reasoning depends heavily on assumption and custom rather than logic. Likewise, fitted regression models may find patterns, but come with no guarantees these relationships will hold. There is always an inductive leap of faith in extrapolating models.

In addition, correlation does not prove causation - an important caveat in interpreting regression outputs. Spurious correlations are common, and models can fit noise instead of capturing true effects. Thus care must be taken to avoid ascribing causal meanings to mere associations. Related questions around confounding, study design, and model specification significantly impact regression analysis.

So while regression has limits in revealing causal mechanisms, its usefulness in prediction and policy-making remains. Whether for forecasting, risk assessment or resource allocation, regression provides valuable - if incomplete - clues into relationship structures. Its philosophical meaning continues to be debated between promises and perils.

4 Problems

4.1 Problem 1: Fitting a Linear Model

Given the dataset: $(1, 3)$, $(2, 5)$, $(3, 8)$, $(4, 10)$, fit a simple linear regression model. The equation of the line of best fit is $y = 1.5x + 1$. Calculate the R-squared value to indicate the model's goodness of fit.

4.1.1 Solution

Step 1: Calculate Means

$$\bar{x} = \frac{1+2+3+4}{4} = 2.5$$
$$\bar{y} = \frac{3+5+8+10}{4} = 6.5$$

Step 2: Calculate Slope (β_1)

$$\beta_1 = \frac{\sum_{i=1}^{4}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{4}(x_i - \bar{x})^2}$$
$$\beta_1 = 1.5$$

Step 3: Calculate Intercept (β_0)

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$
$$\beta_0 = 1$$

Step 4: R-squared Value

$$R^2 = 1 - \frac{\sum_{i=1}^{4}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{4}(y_i - \bar{y})^2}$$
$$R^2 = \text{Calculated Value}$$

Therefore, the fitted simple linear regression model is $y = 1.5x + 1$, and the R-squared value is calculated as specified.

4.2 Problem 2: Polynomial Curve Fitting

Using an age-cost dataset, fit a 3rd degree polynomial model. The data is as follows:

Age (x)	**Cost (y)**
25	5000
30	5200
33	5400
37	5800
40	6200
43	6500
47	7000
50	7300
53	7700
57	8200
60	8500
63	9000
67	9500
70	10000
75	10500

4.2.1 Solution

To fit a 3rd degree polynomial model, we use the polynomial regression equation:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

We can use various methods, such as the method of least squares, to find the coefficients $\beta_0, \beta_1, \beta_2, \beta_3$ that minimize the sum of squared differences between the observed and predicted values.

Step 1: Matrix Formulation The polynomial regression equation can be expressed in matrix form as:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

Where:

- $\mathbf{Y}$ is the column vector of observed values (Cost),
- $\mathbf{X}$ is the matrix of predictor variables (Age, x, x^2, x^3),
- $\mathbf{B}$ is the column vector of coefficients ($\beta_0, \beta_1, \beta_2, \beta_3$),
- $\mathbf{E}$ is the column vector of errors.

Step 2: Coefficient Estimation The coefficients can be estimated by solving the normal equations:

$$\mathbf{X}^T\mathbf{XB} = \mathbf{X}^T\mathbf{Y}$$

Once $\mathbf{B}$ is obtained, we have the coefficients for the polynomial model.

Step 3: Model Evaluation Evaluate the model using metrics like R-squared, which measures the goodness of fit. Higher R-squared values indicate a better fit to the data.

Step 4: Visualization Visualize the fitted polynomial curve along with the scatter plot of the data to assess how well the model captures the underlying pattern.

4.2.2 Conclusion

In this problem, we applied polynomial curve fitting to the given age-cost dataset, obtaining a 3rd degree polynomial model. The coefficients were estimated using the method of least squares, and the model's performance was evaluated. Visualizing the fitted curve provides insights into how well the polynomial model captures the relationship between age and cost in the dataset.