Example 5.4: Let us consider the data given in the Table 5.3 with actual and predicted values. Find standard error estimate.

Solution: The observed value or the predicted value is given below in Table 5.6.

Table 5.6: Sample Data

x_i	y_i	Predicted Value	$(y - \hat{y})^2$
1	1.5	1.46	$(1.5 - 1.46)^2 = 0.0016$
2	2.9	2.02	$(2.9 - 2.02)^2 = 0.7744$
3	2.7	2.58	$(2.7 - 2.58)^2 = 0.0144$
4	3.1	3.14	$(3.1 - 3.14)^2 = 0.0016$

The sum of $(y - \hat{y})^2$ for all $i = 1, 2, 3$ and 4 (i.e., number of samples $n = 4$) is 0.792. The standard deviation error estimate as given in Eq. (5.20) is:

$$\sqrt{\frac{0.792}{4-2}} = \sqrt{0.396} = 0.629$$

5.5 MULTIPLE LINEAR REGRESSION

Multiple regression model involves multiple predictors or independent variables and one dependent variable. This is an extension of the linear regression problem. The basic assumptions of multiple linear regression are that the independent variables are not highly correlated and hence multicollinearity problem does not exist. Also, it is assumed that the residuals are normally distributed.

For example, the multiple regression of two variables x_1 and x_2 is given as follows:

$$\begin{aligned} y &= f(x_1, x_2) \\ &= a_0 + a_1x_1 + a_2x_2 \end{aligned} \tag{5.21}$$

In general, this is given for 'n' independent variables as:

$$\begin{aligned} y &= f(x_1, x_2, x_3, \ldots, x_n) \\ &= a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n + \varepsilon \end{aligned} \tag{5.22}$$

Here, $(x_1, x_2, \ldots, x_n)$ are predictor variables, y is the dependent variable, $(a_0, a_1, \ldots, a_n)$ are the coefficients of the regression equation and ε is the error term. This is illustrated through Example 5.5.

Example 5.5: Apply multiple regression for the values given in Table 5.7 where weekly sales along with sales for products x_1 and x_2 are provided. Use matrix approach for finding multiple regression.

Table 5.7: Sample Data

x_1 (Product One Sales)	x_2 (Product Two Sales)	y Output Weekly Sales (in Thousands)
1	4	1
2	5	6
3	8	8
4	2	12

Solution: Here, the matrices for Y and X are given as follows:

$$X = \begin{pmatrix} 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 8 \\ 1 & 4 & 2 \end{pmatrix} \text{ and } Y = \begin{pmatrix} 1 \\ 6 \\ 8 \\ 12 \end{pmatrix}.$$

The coefficient of the multiple regression equation is given as $a = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}$.

The regression coefficient for multiple regression is calculated the same way as linear regression:

$$\hat{a} = ((X^T X)^{-1} X^T)Y \tag{5.23}$$

Using Eq. (5.23), and substituting the values (Similar to Problem 5.2), one gets $\hat{a}$ as:

$$= \left(\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 4 & 5 & 8 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 8 \\ 1 & 4 & 2 \end{pmatrix} \right)^{-1} \times \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 4 & 5 & 8 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 6 \\ 8 \\ 12 \end{pmatrix}$$

$$= \begin{pmatrix} -1.69 \\ 3.48 \\ -0.05 \end{pmatrix}$$

Here, the coefficients are $a_0 = -1.69$, $a_1 = 3.48$ and $a_2 = -0.05$. Hence, the constructed model is:

$y = -1.69 + 3.48x_1 - 0.05x_2$

5.6 POLYNOMIAL REGRESSION

If the relationship between the independent and dependent variables is not linear, then linear regression cannot be used as it will result in large errors. The problem of non-linear regression can be solved by two methods:

1. Transformation of non-linear data to linear data, so that the linear regression can handle the data
2. Using polynomial regression

Transformations

The first method is called transformation. The trick is to convert non-linear data to linear data that can be handled using the linear regression method. Let us consider an exponential function $y = ae^{bx}$. The transformation can be done by applying log function to both sides to get:

$$\ln y = bx + \ln a \tag{5.24}$$

Similarly, power function of the form ($y = ax^b$) can be transformed by applying log function on both sides as follows:

$$\log_{10} y = b\log_{10} x + \log_{10} a \tag{5.25}$$

Once the transformation is carried out, linear regression can be performed and after the results are obtained, the inverse functions can be applied to get the desire result.

Polynomial Regression

It can handle non-linear relationships among variables by using n^{th} degree of a polynomial. Instead of applying transforms, polynomial regression can be directly used to deal with different levels of curvilinearity.

Polynomial regression provides a non-linear curve such as quadratic and cubic. For example, the second-degree transformation (called quadratic transformation) is given as: $y = a_0 + a_1x + a_2x^2$ and the third-degree polynomial is called cubic transformation given as: $y = a_0 + a_1x + a_2x^2 + a_3x^3$. Generally, polynomials of maximum degree 4 are used, as higher order polynomials take some strange shapes and make the curve more flexible. It leads to a situation of overfitting and hence is avoided.

Let us consider a polynomial of 2^{nd} degree. Given points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the objective is to fit a polynomial of degree 2. The polynomial of degree 2 is given as:

$$y = a_0 + a_1x + a_2x^2 \tag{5.26}$$

Such that the error $E = \sum_{i=1}^{n}[y_i - (a_0 + a_1x_i + a_2x_i^2)]^2$ is minimized. The coefficients a_0, a_1, a_2 of Eq. (5.26) can be obtained by taking partial derivatives with respect to each of the coefficients as $\frac{\partial E}{\partial a_1}, \frac{\partial E}{\partial a_1}, \frac{\partial E}{\partial a_2}$ and substituting it with zero. This results in 2 + 1 equations given as follows:

$$na_0 + \left(\sum_{i=1}^{n} x_i\right)a_1 + \left(\sum_{i=1}^{n} x_i^2\right)a_2 = \sum_{i=1}^{n} y_i$$

$$\left(\sum_{i=1}^{n} x_i\right)a_0 + \left(\sum_{i=1}^{n} x_i^2\right)a_1 + \left(\sum_{i=1}^{n} x_i^3\right)a_2 = \sum_{i=1}^{n} x_iy_i$$

$$\left(\sum_{i=1}^{n} x_i^2\right)a_0 + \left(\sum_{i=1}^{n} x_i^3\right)a_1 + \left(\sum_{i=1}^{n} x_i^4\right)a_2 = \sum_{i=1}^{n} x_i^2y_i \tag{5.27}$$

The best line is the line that minimizes the error between line and data points. Arranging the coefficients of the above equation in the matrix form results in:

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix}\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum(x_i, y_i) \\ \sum(x_i^2, y_i) \end{bmatrix} \tag{5.28}$$

This is of the form $Xa = B$. One can solve this equation for a as:

$$a = X^{-1}B \tag{5.29}$$

Example 5.6: Consider the data provided in Table 5.8 and fit it using the second-order polynomial.

1	1	1	1	1	1	1
2	4	8	4	16	8	16
3	9	27	9	81	27	81
4	15	60	16	240	64	256
$\Sigma x_i = 10$	$\Sigma y_i = 29$	$\Sigma x_i y_i = 96$	$\Sigma x_i^2 = 30$	$\Sigma x_i^2 y_i = 338$	$\Sigma x_i^3 = 100$	$\Sigma x_i^4 = 354$

It can be noted that, $N = 4$, $\Sigma y_i = 29$, $\Sigma x_i y_i = 96$, $\Sigma x_i^2 y_i = 338$. When the order is 2, the ma[...] using Eq. (5.28) is given as follows:

$$\begin{bmatrix} 4 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 29 \\ 96 \\ 338 \end{bmatrix}$$

Therefore, using Eq. (5.29), one can get coefficients as:

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 4 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{bmatrix}^{-1} \times \begin{bmatrix} 29 \\ 96 \\ 338 \end{bmatrix} = \begin{pmatrix} -0.75 \\ 0.95 \\ 0.75 \end{pmatrix}.$$

This leads to the regression equation using Eq. (5.26) as:

$$y = -0.75 + 0.95\,x + 0.75\,x^2$$

5.7 LOGISTIC REGRESSION

Linear regression predicts the numerical response but is not suitable for predicting the categor[...] variables. When categorical variables are involved, it is called classification problem. Logi[...] regression is suitable for binary classification problem. Here, the output is often a categor[...] variable. For example, the following scenarios are instances of predicting categorical variables.

1. Is the mail spam or not spam? The answer is yes or no. Thus, categorical depend[...] variable is a binary response of yes or no.
2. If the student should be admitted or not is based on entrance examination ma[...] Here, categorical variable response is admitted or not.
3. The student being pass or fail is based on marks secured.

Thus, logistic regression is used as a binary classifier and works by predicting the probability of the categorical variable. In general, it takes one or more features x and predicts the response y. If the probability is predicted via linear regression, it is given as:

$$p(x) = a_0 + a_1 x$$

Hence, logistic regression tries to model the probability of the particular response variable. In email classification problem, say normal email or spam, if the probability of the response variable is 0.7, then there is a 70% possibility of a normal mail.

Linear regression generated value is in the range $-\infty$ to $+\infty$, whereas the probability of the response variable ranges between 0 and 1. Hence, there must be a mapping function to map the value $-\infty$ to $+\infty$ to 0–1. The core of the mapping function in logistic regression method is sigmoidal function. A sigmoidal function is a '*S*' shaped function that yields values between 0 and 1. This is known as logit function. This is mathematically represented as:

$$\text{logit}(x) = \frac{1}{1 + e^{-x}} \tag{5.30}$$

Here, x is the independent variable and e is the Euler number. The purpose of the logit function is to map any real number to zero or 1.

Logistic regression can be viewed as an extension of linear regression, but the only difference is that the output of linear regression can be an extremely high number. This needs to be mapped into the range 0–1, as probability can have values only in the range 0–1. This problem is solved using log odd or logit functions. What is the difference between odds and probability? Odds and probability (or likelihood) are two sides of a coin and represent uncertainty. The odds are defined as the ratio of the probability of an event and probability of an event that is not happening. This is given as:

$$odd = \frac{\textit{probability} \text{ of an event}}{\textit{probability} \text{ of an non} - \text{event}} = \frac{p}{1 - p} \tag{5.31}$$

Log-odds can be taken for the odds, resulting in:

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = a_0 + a_1 x \tag{5.32}$$

Here, log(.) is a logit function or log odds function. One can solve for $p(x)$ by taking the inverse of the above function as:

$$p(x) = \frac{\exp(a_0 + a_1 x)}{1 + \exp(a_0 + a_1 x)} \tag{5.33}$$

This is the same sigmoidal function. It always gives the value in the range 0–1. Dividing the numerator and denominator by the numerator, one gets:

$$p(x) = \frac{1}{1 + \exp(-a_0 - a_1 x)} \tag{5.34}$$

One can rearrange this by taking the minus sign outside to get the following logistic function:

$$p(x) = \frac{1}{1 + \exp(-(a_0 + a_1 x))} \tag{5.35}$$

Here, x is the explanatory or predictor variable, e is the Euler number, and a_0, a_1 are regression coefficients. The coefficients a_0, a_1 can be learned and the predictor predicts $p(x)$ dire… using the threshold function as:

$$y = \begin{cases} 1 & \text{if } p(x) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

(5.…

Example 5.7: Let us assume a binomial logistic regression problem where the classes are pass a… fail. The student dataset has entrance mark based on the historic data of those who are selec… or not selected. Based on the logistic regression, the values of the learnt parameters are $a_0 = 1$ a… $a_1 = 8$. Assuming marks of $x = 60$, compute the resultant class.

Solution: The values of regression coefficients are $a_0 = 1$ and $a_1 = 8$, and given that $x = 60$.

Based on the regression coefficients, z can be computed as:

$$z = a_0 + a_1 x$$
$$= 1 + 8 \times 60 = 481$$

One can fit this in a sigmoidal function using Eq. (5.30) to get the probability as:

$$y = \frac{1}{1 + \exp(-481)} = \frac{1}{2.271} = 0.44$$

If we assume the threshold value as 0.5, then it is observed that $0.44 < 0.5$, therefore, t… candidate with marks 60 is not selected.

To determine the relationship between dependant and independent variables, parameters nee… to be obtained. In logistic regression, the parameters are obtained through maximum likeliho… function (MLE) using the training data. The aim is to learn the values of parameters of the logist… model (a's) by minimizing the error in the probability predicted by the model.

There can be many different sets of coefficients available. The optimal value of the paramete… is obtained by using the MLE function, which is a set of coefficients for which the probability… getting the observed data is maximum.

If π is the success of the outcome and $1 - \pi$ is the failure of the outcome, then the likelihoo… function is given as:

$$L(a : y) = \prod_{i=1}^{n} \left(\frac{\pi_i}{1 - \pi_i}\right)^{y_i} (1 - \pi_i) \quad (5.3…$$

To determine the value of the parameters, the log of the likelihood function is taken. Techniqu… like Newton method can be used to maximize the log-likelihood of the function.

Logistic regression is suitable for binary classification. The idea can be extended for multip… classes called multinomial logistic regression. Let us assume that there are three classes 1, 2 and… Then, the multinomial logistic regression creates three classification problems – class 1 and Not cla… 1, class 2 and Not class 2, and finally class 3 and Not class 3. Three problems are simultaneous… used to find the maximum probability relative to others to get the appropriate class.

Logistic regression is a simple and efficient method for binary classification. The model ca… be easily interpreted too. The disadvantages of logistic regression are that multinomial logist… regression cannot handle many attributes and can handle only linear features. Also, if all th… attributes have multicollinearity problem, the logistic method does not work effectively.

5.8 RIDGE, LASSO, AND ELASTIC NET REGRESSION

Recollect that Ordinary Least Square (OLS) fits a line for the data points that minimize the sum of squared error between the data points and the line of fit.

There are two issues that need to be considered before discussing regularization methods. One is bias and another is variance.

1. Bias – If the selected model does not fit the dataset well, then this error is called bias.
2. Variance – If the model works nicely for the trained data but is not representative for the entire universe of the possible data, it is called variance.

Consider the Figure 5.6 where the errors of the individual points are shown.

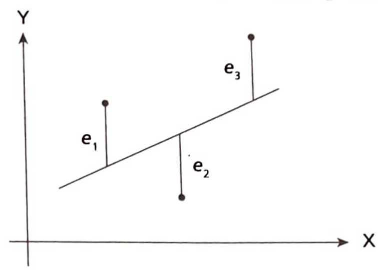

Figure 5.6: The Variance of the Model

The vertical line that measures the error between the data points and the line of fit indicates the variance. The sum of the individual error squares refers to the amount of variance that is not captured by the model. It should be minimal.

If the line fits the training points but lacks the ability to fit the testing phase points, it indicates generalization error. A line of good fit should have generalization capability and should have the least error for the training as well as testing phase. A line that minimizes the error for the training points and more error for the testing phase is called overfitting problem. Overfitting gives more variance.

The main idea to solve variance is to introduce a small bias to create a ridge line to reduce variance. Consider the following Figure 5.7. Here, the first two points are training data points and the next two points are test data points. If a line is fit through the training data, then it may not work properly for the test data points. In other words, the model lacks generality.

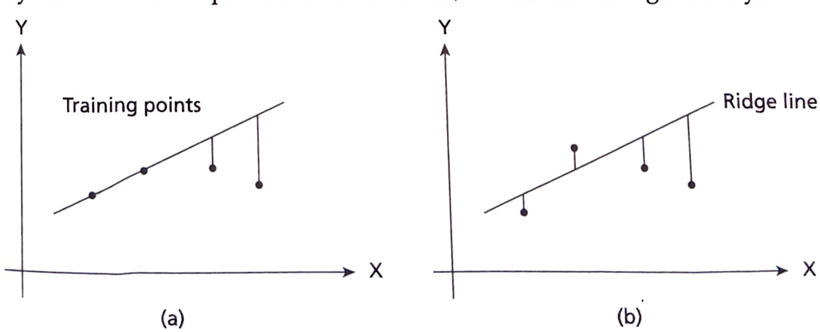

Figure 5.7: (a) OLS Fit Line (b) Ridge Line

The solution for this problem is to fit a ridge line by changing the slope of the OLS line. This introduces a bias in the model and because of the bias, even though this line does not fit the training data, the overall variance gets reduced. So, the aim of regularization is to reduce the test data error.

Normally, low variance algorithm has high bis and vice versa. If we have high variance and low bias, the problem is called underfitting. The reciprocal problem is called overfitting. So, there must be a balance between bias and variance. This can be achieved by removing the unnecessary attributes of the dataset. This is done by two ways:

1. Feature selection algorithms remove unnecessary attributes. The forward selection and backward selection algorithms and principle component analysis are discussed in Chapter 2.
2. Another way to solve the problem is to use regularization methods that add a penalty to penalize the regression coefficients to simplify the model.

Let us discuss about regularization methods now.

5.8.1 Ridge Regularization

Ridge regression is used to create a parsimonious model. It is useful when there are many attributes that exceed the number of observations with strong correlation among them. Ridge regression is a plain, linear regression model whose regression coefficients are not estimated by OLS but by a ridge estimator. Ridge estimator is a biased estimator that has lower variance than the OLS estimator. The mean square error of the ridge estimator (sum of the variance and square of its bias) is lower than the estimator of OLS method. The estimator for ridge regression is given as follows:

$$\text{Sum of squared residuals} + \lambda \times slope^2$$

That is, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \times slope^2$ (5.38)

Here, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the sum of the squared residuals and $\lambda \times slope^2$ is the penalty term added to change the slope of the line of fit to reduce variance. The purpose of the penalty is to penalize the large regression coefficients. If the value of the regression is more, then penalty is more. This is done by reducing the variance by adding a small bias. The change of slope is the main aspect of ridge regression. It reduces the slope of the line. If the slope of the line is small, then the change in prediction is barely noticeable. Thus, ridge line creates a small slope and is hence insensitive.

Ridge regression uses L_2 regularization. This limits the size of the coefficients using a penalty function. L_2 Penalty equals the square of the magnitude of all coefficients. The penalty is controlled by a parameter λ. When $\lambda = 0$, ridge regression is same as OLS. When λ is infinity, the slope becomes almost zero and hence shrinks all coefficients to 0. Thus, ridge regression avoids the problem of overfitting by shrinking the less important variables closer to zero. The choice of λ is very crucial. It is obtained by leave-one-out validation method. The details of this validation method are discussed in Chapter 3. The procedure for selecting λ is given as follows:

1. Choose a possible set of values of the penalty
2. Exclude the N^{th} observed data and compute the penalty

3. Compute the penalty out of samples that are excluded
4. Compute the mean square error and pick the penalty that minimizes the MSE

The matrix formulation of ridge regression is given as follows:

$$\hat{a} = (X'X + kI)^{-1}X'Y \tag{5.39}$$

The major problem is the selection of 'k' in Eq. (5.39). Ridge regression is also useful for logistic regression where it is expressed as sum of likelihoods + $\lambda \times slope^2$. This idea can be extended for many independent variables' regression coefficients except y-intercept.

5.8.2 LASSO

LASSO regression is better than ridge regression. LASSO stands for 'Least Absolute Shrinkage and Selection Operator'. This uses L_1 regularization.

This adds a penalty factor that equals the square of the magnitude of all coefficients. This penalty results in a simple model given as:

$$\text{Sum of squared residuals} + \lambda \times |slope|$$

Thus, LASSO model is given as:

$$\hat{a} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=0}^{k}|a_i| \tag{5.40}$$

By controlling the factor λ, where the high values of λ force many coefficients to zero, LASSO performs both shrinkage and variable selection.

The difference between Ridge and LASSO regression methods is given in the following Table 5.10.

Table 5.10: Difference between Ridge and LASSO Regression

Ridge Regression	LASSO Regression
Penalty term is $\lambda \times slope^2$	Penalty term is $\lambda \times |slope|$
This method shrinks the regression coefficients of less important variables closer to zero.	This method shrinks the regression coefficients of less important variables to zero giving a compact model.
Not useful for feature selection	Good feature selector by removing all irrelevant variables

5.8.3 Elastic Net

Elastic Net is a hybrid method of combining both Ridge and LASSO regression methods. The Elastic Net is given as follows:

$$\text{Sum of squared residuals} + \lambda_1 + \underbrace{|v_1| + |v_2| + \cdots + |v_k|}_{Lasso} + \lambda_2 \times \underbrace{|v_1|^2 + |v_2|^2 + \cdots + |v_k|^2}_{Ridge} \tag{5.41}$$

Here, $v_1, v_2, \ldots, v_k$ are dependent variables of the regression method.

It uses separate penalty factors for Lasso and Ridge regression methods. When λ_1 and λ_2 are zero, then Elastic Net reduces to simple OLS method. When $\lambda_1 = 0$, Elastic Net serves as ridge regression and when $\lambda_2 = 0$, Elastic Net serves as LASSO regression. When both λ_1 and λ_2 are greater than zero, then it serves as hybrid technique.

Elastic Net groups the variables and shrinks the parameters associated with the correlated variables and makes it closer to zero or zero, thereby removing the irrelevant attributes. Elastic Ne is good in situations where multicollinearity problem exists among the independent variables o the regression method.

Summary

1. Regression is a supervised learning method that can predict continuous variables.
2. Regression analysis is used to model the relationship between one or more independent variable and a dependent variable whereas in multiple regression problems, the output is a combination of predictor variables.
3. Regression is used for prediction and forecasting. This determines the change in response variable when one exploration variable is varied while keeping all other parameters constant. This is used to determine the relationship each of the exploratory variables exhibits.
4. Scatter plot is a plot of explanatory variable and response variable. It is a 2D graph showing the relationship between two variables. The quality of the regression analysis is determined by the factors of correlation and causation.
5. Linear regression model can be created by fitting a line among the scattered data points. The line is of the form: $y = a_0 + a_1x$. Here, a_0 is the intercept which represents the bias and a_1 represents the slope of the line.
6. Multiple regression model is an extension of linear regression model that involves multiple predic tors. It is also used to find relationships among variables, important features and construct models
7. Polynomial regression can handle non-linear relationships among variables by using degrees polynomial to make non-regression line. If the relationship is not linear, then polynomial regressi can be used at different levels of curvilinearity.
8. Logistic regression is suitable for binary classification problem. Here, the output is often a bin variable. For example, consider a question – Is the mail spam or not spam? The answer is yes or Thus, logistic regression is used to predict categorical dependant variable.
9. Ridge regression is used to create a parsimonious model by introducing ridge estimator that in duces some bias in the model to reduce variance.
10. LASSO stands for "Least Absolute Shrinkage and Selection Operator". This uses L_1 regularizatio remove irrelevant variables.

Key Terms

- **Regression Analysis** – A mathematical technique of modelling the relationship between input output variables.
- **Independent Variable** – Th