Regression Analysis

Rohan Sinha(230041030)
Maanas(230041020)
L Praveen Kumar(230041018)

April 10, 2025

Contents

1 Introduction

1.1 Importance of Regression

Regression analysis is one of the most widely used techniques for analysing multiple factor data. Its usefulness arises from its logical process of using an equation to generate relationship between a variable of interest(response variable) and set of related values(regressor).

1.2 Objective

Prediction of various things such as house prices, medical expenses etc has always been a key topic of interest in Regression Analysis. Today, we will be using a multiple regression model to accurately predict house prices based on various factors.

1.3 Scope

We will be using 4 features- Square footing(in feet), number of bedrooms, location and age(in yrs) of property to predict the house prices on a dataset of 20 points. Also we will be assuming that the model is linear based on the scope of our project and the estimator for our model will be Least Square Estimator.

2 Data Description

2.1 Data

SqFt	Bedrooms	Location	Age	Price
2000	3	Urban	5	450
1500	2	Urban	10	300
1800	3	Urban	8	400
1200	2	Rural	15	180
2200	4	Urban	3	420
1600	3	Urban	12	350
1400	2	Rural	20	170
2500	4	Urban	2	500
1700	3	Urban	7	360
1300	2	Rural	18	160
1900	3	Urban	6	380
2100	4	Urban	4	460
1550	3	Urban	11	320
1350	2	Rural	16	175
2300	4	Urban	1	490
1450	2	Urban	9	290
1750	3	Rural	14	240
2400	4	Urban	3	480
1600	2	Urban	10	310
1250	2	Rural	22	150

2.2 Feature Description

1. Square Footing- The square footage of a house refers to the total interior area of the house. It is calculated by sketching a floor plan of the interior and then breaking it down to measurable rectangles.
2. Bedrooms- It is basically the number of bedrooms in the house.
3. Location- It refers to the location of the house where it is located. If the property is in cities then it is said to be in urban else if the property is in a village then it it said to be in rural area.
4. Age- As the name suggests, this refers to the time it was built/constructed to the time till the present.

2.3 Data Preprocessing

Since all the features except location are numerical, we will be using binary assignment to make it numerical.

Location	IsUrban	IsRural
Rural	0	1
Urban	1	0

3 Methodology

3.1 Regression Model

A multiple linear regression model is defined as

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i$ where ε has mean 0 and some variance σ^2 for $1 \leq i \leq 20$

Since we have 5 parameters, we will only take upto β_5. Here, y is the response variable which denotes the price(in dollars), and x_i are the regressor variables where x_1 is the square footing, x_2 is the number of bedrooms, x_3 is the boolean value of IsUrban, x_4 is the age of property and x_5 is the boolean value of IsRural. Now, we can see that there is a linear relation between IsRural and IsUrban variable and thus we can drop one of them from the table. Let us drop IsRural variable. Thus equation becomes.

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon$ for $1 \leq i \leq 20$

Since there are 20 points, it is easier to write the equation in matrix form as

$y = X\beta + \varepsilon$ Here X is the feature matrix given as

$$\mathbf{X} = \begin{bmatrix} 1 & 2000 & 3 & 1 & 5 \\ 1 & 1500 & 2 & 1 & 10 \\ 1 & 1800 & 3 & 1 & 8 \\ 1 & 1200 & 2 & 0 & 15 \\ 1 & 2200 & 4 & 1 & 3 \\ 1 & 1600 & 3 & 1 & 12 \\ 1 & 1400 & 2 & 0 & 20 \\ 1 & 2500 & 4 & 1 & 2 \\ 1 & 1700 & 3 & 1 & 7 \\ 1 & 1300 & 2 & 0 & 18 \\ 1 & 1900 & 3 & 1 & 6 \\ 1 & 2100 & 4 & 1 & 4 \\ 1 & 1550 & 3 & 1 & 11 \\ 1 & 1350 & 2 & 0 & 16 \\ 1 & 2300 & 4 & 1 & 1 \\ 1 & 1450 & 2 & 1 & 9 \\ 1 & 1750 & 3 & 0 & 14 \\ 1 & 2400 & 4 & 1 & 3 \\ 1 & 1600 & 2 & 1 & 10 \\ 1 & 1250 & 2 & 0 & 22 \end{bmatrix}$$

Now for calculating the estimator of β given by the equation $haty = X\hat{\beta}$, we will use the formula-
$\hat{\beta} = (X^TX)^{-1}X^Ty$

3.2 Code

```python
# -*- coding: utf-8 -*-
"""RegBeta.ipynb

Automatically generated by Colab.

Original file is located at
    https://colab.research.google.com/drive/1
        zNCrIo6rvHQRRROcsLQmEegHDhgquXcn
"""

import numpy as np
import pandas as pd
```

```

from google.colab import files
uploaded = files.upload()

df = pd.read_csv("feature_matrix_no_intercept.csv")
y = np.array([
    450, 300, 400, 180, 420, 350, 170, 500, 360, 160,
    380, 460, 320, 175, 490, 290, 240, 480, 310, 150
])
X = df.values
X = df.drop(columns=["Rural"]).values
X = np.hstack((np.ones((X.shape[0], 1)), X))

XtX = X.T @ X
XtX_inv = np.linalg.inv(XtX)
XtY = X.T @ y
beta = XtX_inv @ XtY

# Step 7: Output the coefficients
print("Regression coefficients (Beta):")
for i, b in enumerate(beta):
    print(f"Beta{i}: {b:.4f}")
```

4 Results

4.1 Regression Coefficients

The regression coefficients calculated from the code are-

Coefficient	Value
Intercept (β_0)	28.23
Square Footage (β_1)	0.12
Bedrooms (β_2)	21.79
Urban (β_3)	91.47
Age (β_4)	-3.51

Table 1: Regression coefficients for house price prediction

The regression equation we get is-
$\hat{y} = 28.23 + 0.12x_1 + 21.79x_2 + 91.47x_3 - 3.51x_4$

4.2 Residual

Actual (y)	Predicted ($\hat{y}$)	Residual ($y - \hat{y}$)
450	407.73	42.27
300	308.34	-8.34
400	373.18	26.82
180	163.29	16.71
420	460.56	-40.56
350	335.13	14.87
170	169.76	0.24
500	500.10	-0.10
360	364.68	-4.68
160	164.77	-4.77
380	392.21	-12.21
460	445.04	14.96
320	332.63	-12.63
175	177.80	-2.80
490	479.59	10.41
290	305.85	-15.85
240	254.65	-14.65
480	484.58	-4.58
310	320.35	-10.35
150	144.73	5.27

Table 2: Actual vs Predicted Values and Residuals

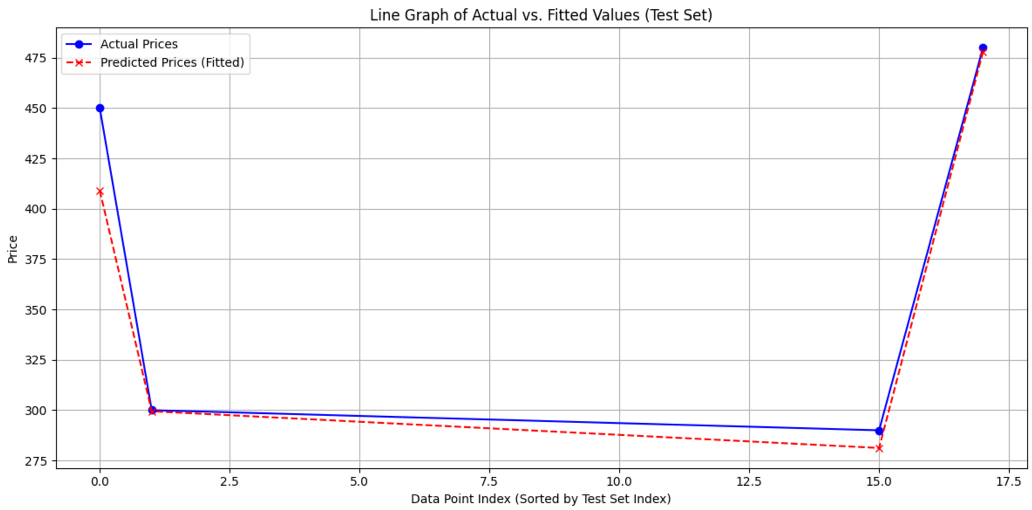

Figure 1: Linear regression visualization

This line graph compares the actual (blue) and predicted (red dashed) prices from a regression model across the test dataset. The closer the red line follows the blue line, the better the model's predictions. Here, the model captures the general trend of price changes but shows noticeable deviations at certain data points, indicating prediction errors. The x-axis represents the sorted order of data points in the test set, while the y-axis displays the price.

4.3 Interpretation of Coefficients

In the equation, we can see that $\beta_1, \beta_2, \beta_3$ are positive while β_4 is negative. This implies the price of house has a direct relation with square footing, number of bedrooms and location being Urban. This is correct as more square footing means more spacious house, more number of bedrooms being a positive factor and location being Urban is more preferable. On the other hand, the age of property is a negative factor as more old the house gets the cheaper it becomes.

4.4 Estimation of σ^2

For estimation of σ^2 as say $\hat{\sigma}^2$, we will use the formula.
$\hat{\sigma}^2 = SS_{Res}/(n-p)$ where $SS_{Res} = \sum_{i=1}^{20}(y_i - \hat{y}_i)^2$
The value of SS_{Res} from the table comes out to be 6969.78 approximately.

Thus $\hat{\sigma}^2 = 6969.78/(20-5) = 464.65$
The standard error is approximately 20.06.

5 References

1. Montgomery, D.C, Peck, E.A, Vining, G.G.(2012). Introduction to linear regression analysis (5th ed.). Wiley.
2. Kaggle, "House Price Prediction Dataset." Retrieved April 10,2025 ,from https://www.kaggle.com/datasets/zafarali27/house-price-prediction-dataset