Multi Linear Regression on Housing Data and Stock Market

Jean Batista

Juniata College

Linear Algebra

Catherine Stenson

3/8/2025

Introduction to Linear Regression

Linear regression is a common technique used to predict a dependent variable (y) using one or more independent variables. In its simplest form simple linear regression, the method examines the relationship between two variables. For instance, if you want to understand how the age of a house affects its price, you would analyze data where the price is the dependent variable and the house's age is the independent variable. In many real-world situations, a single variable cannot fully explain an outcome. A house's price, for example, is influenced not only by its age but also by factors such as living area, lot size, and other features. Multivariable linear regression addresses this complexity by incorporating multiple predictors into the model. Here is the general equation[1]:

$$Y= \beta 0+ \beta 1X1+ \beta 2X2+ \beta 3X3+\ldots+\varepsilon$$

Here, X_1, X_2, X_3, etc. represent the independent variables (like Living Area, Lot Size, Age), β_1, β_2, β_3, etc. are the coefficients that quantify the impact of each variable on the dependent variable (Price), and ε captures the variation not explained by the model.

Data and Context

This technique is ideal for quantitative data where the dependent variable is continuous, such as housing prices. The independent variables can be continuous, categorical (with proper encoding), or a mix of both. Multivariable linear regression is particularly useful when you want to: 1. Understand how several factors simultaneously influence an outcome, 2. Control for confounding variables, and 3. Make predictions as well as draw inferences about relationships between variables. Schneider explains that researchers can determine an appropriate sample size by considering their expected values for the coefficient of determination (r^2) and the regression

coefficient (b). Typically, the study should include at least twenty times as many observations as there are independent variables, in other words, if you are examining two predictors, you should aim for a minimum of 40 observations [2]. For the model to be reliable, the relationship between the dependent and independent variables should be approximately linear, and you should have enough observations relative to the number of predictors, typically at least 20 observations per predictor.

How Multivariable Linear Regression Works with Linear Algebra

The primary method used to estimate the coefficients β0, β1,… in a multivariable linear regression model is Ordinary Least Squares (OLS). OLS aims to minimize the sum of the squared differences between the observed values and the values predicted by the model. In a multivariable context, this process involves more complex math, as the model is working in a multi-dimensional space.

To explain how linear algebra fits into this, we first organize the data into a matrix form. In this case, we have our predictors (Living Area, Lot Size, and Age) represented as vectors in a matrix X, and the dependent variable (Price) as a vector y. Our predictors have the coefficient $\boldsymbol{\beta}$ which we aim to solve using linear algebra. We can achieve this using the normal equation [3]

where $A^T AX = A^T b$

David Austin in proposition 3.1.4 states that if you have an invertible matrix, then multiplying both sides of the equation by its inverse will isolate the unknown vector. Such that if

$Ax = b$ then $A^{-1}b = x$. [4] Using linear algebra, we can solve for β using the formula:

$$\boldsymbol{\beta} = (A^T A)^{-1} A^T b$$

Here's how the linear algebra works:

A^T is the transpose of matrix A, swapping the rows and columns. $A^T A$ is the normal equation that combines the predictors to form a square matrix. The inverse of $A^T A$, denoted $(A^T A)^{-1}$, is computed to "undo" the multiplication and ensure the solution is valid. Finally, multiplying the inverse by $A^T$and b gives the estimated coefficients β.

These coefficients represent the "best-fit" line in a multi-dimensional space, where the line minimizes the error between the predicted and actual house prices. However, if the predictors in the matrix A are highly correlated (some features are similar to each other), the matrix $A^T A$ can become unstable. This is explained further in Draper and Smith, if a matrix has high correlation between columns, the matrix is close to singular and its inverse may not exist or might give unreliable estimates[5]. The least squares method will not produce unique solutions but multiple possible estimates. Meaning the data could be insufficient for the model, or when the model is too complex relative to the available data. To resolve this, more data is needed, or the model should be simplified to better fit the available data.

By using linear algebra to solve the OLS problem, we gain a clear understanding of how each predictor (Living Area, Lot Size, and Age) influences the house price. However, the quality of the model depends heavily on the quality and stability of the data used. For example:

Houses	Living Area(sq.ft)	Lot Size(acres)	Age(years)	Price(USD)
8	1464	0.11	87	108794
9	1216	0.61	101	68353
10	1632	0.23	14	123266

11	2270	4.05	9	309808
12	1804	0.43	0	157946
13	1600	0.36	16	80248
14	1460	0.18	17	135708
15	1548	0.36	0	173723

(At a glance, one might infer that larger living areas and lot sizes generally lead to higher prices, while older houses might be priced lower. However, with just eight observations, any conclusions must be drawn with caution.

Solution

To solve the solution to the regression model we take the data and try to solve the normal equations $A^TAX = A^Tb$. Given our data set we can construct a matrix and vector as such:

```
A = Matrix([
    [1, 1464, 0.11, 87],
    [1, 1216, 0.61, 101],
    [1, 1632, 0.23, 14],
    [1, 2270, 4.05, 9],
    [1, 1804, 0.43, 0],
    [1, 1600, 0.36, 16],
    [1, 1460, 0.18, 17],
    [1, 1548, 0.36, 0]
])
```

In this matrix, we applied the intercept b0 to the first column of the matrix to form our equation:

Y(vector) = matrix[β0 + β1X1(Living Area(sq.ft))+β2 X2(Lot Size(acres))+β3X3(Age(years))]

```
b = vector([108794, 68353, 123266, 309808, 157946, 80248, 135708, 173723])
```

```
AtA = A.transpose() * A
Atb = A.transpose() * b
```

As we isolate the x variable. Remembering the inverse property such that if $Ax = b$, then $A^{-1}b = x$. We use this property to solve for x as $(A^TA)^{-1} * A^Tb = x$.

```
coeff = AtA.inverse() * Atb
```

```
print("The least squares regression coefficients are:")
print("beta_0 (intercept)   =", coeff[0])
print("beta_1 (Living Area) =", coeff[1])
print("beta_2 (Lot Size)    =", coeff[2])
print("beta_3 (Age)         =", coeff[3])
```

```
The least squares regression coefficients are:
beta_0 (intercept)   = 40929.5245332102
beta_1 (Living Area) = 55.5138081162741
beta_2 (Lot Size)    = 35608.3209810648
beta_3 (Age)         = -476.792175158557
```

Thus Price = 40929.52 + 55.51(Living Area) + 35608.32(Lot Size) − 476.79(Age)

Now for calculating R^2

In [5]:

```
# Coefficients
beta_0 = 40929.5245332102  # Intercept
beta_1 = 55.5138081162741  # Living Area coefficient
beta_2 = 35608.3209810648  # Lot Size coefficient
beta_3 = -476.792175158557 # Age coefficient

#House data
data = [
    (1464, 0.11, 87, 108794),
    (1216, 0.61, 101, 68353),
    (1632, 0.23, 14, 123266),
    (2270, 4.05, 9, 309808),
    (1804, 0.43, 0, 157946),
    (1600, 0.36, 16, 80248),
    (1460, 0.18, 17, 135708)
]

# Extract the observed values and actual price
X = [entry[:-1] for entry in data]  # Living Area, Lot Size, Age
y_actual = [entry[-1] for entry in data]  # Actual Price

# Compute the predicted values using the regression equation
y_predicted = [beta_0 + beta_1 * x[0] + beta_2 * x[1] + beta_3 * x[2] for x in X]

# Compute the mean of the actual prices
y_mean = sum(y_actual) / len(y_actual)

# Compute SST (Total Sum of Squares)
SST = sum((yi - y_mean) ** 2 for yi in y_actual)

# Compute SSE (Residual Sum of Squares)
SSE = sum((yi - y_pred) ** 2 for yi, y_pred in zip(y_actual, y_predicted))

# Compute R^2
R_squared = 1 - (SSE / SST)

# Display the result
R_squared
```

Out[5]: 0.895075020156838

The Results

The regression output for the housing dataset shows that about 89% of the variation in house prices can be explained by the model. Essentially, multivariable linear regression tries to find the "best-fit" line that represents the relationship between factors like living area, lot size, and the age of a house with its price. It does this by using math to minimize the difference between the actual prices and the prices predicted by the model.

In our case, we computed $\boldsymbol{R^2}$, which is a measure of how well the model's predictions match the actual data. An $\boldsymbol{R^2}$ of 0.89 means the model does a good job of explaining house prices based on the features we provided.

However, this process involves some complex math, and if the data used to build the model isn't great, like if the variables are too similar or there's not enough data, the math can become unstable.

Uses cases in other scenarios

A similar approach can be applied to predict stock prices. In one example[6], scholars wrote an article on a multivariable linear regression model that forecasts a stock's daily high price based on two factors: the stock's opening price and the NASDAQ's opening price. Here, the stock's opening price is the price at which the stock first trades when the market opens, reflecting the initial investor sentiment and overnight news, while the daily high price is the maximum price the stock reaches during trading hours, capturing its intraday volatility. Predicting the daily high price is significant because it offers insights into potential profit opportunities, helps investors determine optimal entry and exit points, and supports better risk management during trading. To build a reliable model, historical stock data is extracted using the Yahoo Finance library, which

provides current and comprehensive market information. This data is then divided into a training period of one year and a testing period of one month to ensure that the model is well-calibrated and can generalize effectively. The model for Apple's stock is:

Apple's predicted high price = -0.2514 + 0.9965(Apple open) + 0.0230(NASDAQ open) + ε

Despite achieving a high r^2 of 0.91 and a low root mean square error of 1.14, the model's simplicity is a key limitation. It uses only two predictors, the stock's daily opening price and the NASDAQ's opening price, which capture only basic market conditions. As Schneider points out, when a model relies on just two independent variables, it should ideally be supported by at least 40 observations to yield reliable estimates; any shortfall in sample size can further undermine the model's robustness. In reality, many other factors such as investor sentiment, economic indicators, trading volumes, technical patterns, and unexpected events like economic shifts or geopolitical tensions can significantly influence stock prices. Moreover, if the training data comes from a period of unusual calm or volatility, the model might not perform well when applied to different time periods. This narrow focus means the model might be underfit by missing key dynamics, or even overfit if the specific time frame isn't representative of broader market conditions.

Both the housing and stock models demonstrate that a strong statistical fit on paper does not always translate into reliable predictions in practice. The housing model, based on only eight observations and three predictors, faces issues like large standard errors and potential multicollinearity. Similarly, the stock model's reliance on just two predictors overlooks important influences on stock prices. In both cases, it's crucial to balance model simplicity with sufficient data and relevant predictors to capture the true complexity of the underlying phenomena.

Conclusion

Multivariable linear regression is a powerful tool that leverages linear algebra to combine multiple predictors for explaining and forecasting outcomes. Whether predicting house prices or stock market trends. This paper has shown how starting with a simple linear relationship can evolve into a more robust multivariable approach that accounts for several influencing factors. We explored how the method works by minimizing the errors between observed and predicted values using Ordinary Least Squares, and how linear algebra helps solve for the best-fitting coefficients.

Throughout the discussion, examples like the housing dataset and stock price prediction underscored important lessons. With the housing data, we saw that even a strong overall fit can be misleading when based on limited observations and a small number of predictors. The stock example further highlighted that while a model may show promising statistics on paper, relying on only a couple of variables might overlook the many unpredictable factors driving market behavior. These cases remind us that both overfitting and underfitting are real concerns; a model must balance simplicity with the inclusion of enough relevant data to truly capture real-world complexity.

In essence, a solid multivariable linear regression model requires more than just a good statistical fit. It depends on sufficient data, thoughtful selection of predictors, and careful interpretation of results. When these elements are in harmony, the model not only provides statistical insights but also offers meaningful, real-world predictions.

Citations :

[1]Huang Y, 2019, Multiple Linear Regression Handouts, Dep of Statistics, University of Chicago, https://www.stat.uchicago.edu/~yibi/teaching/stat222/2019/MLR_2019.pdf

[2] Schneider A, Hommel G, Blettner M. 2010 Linear regression analysis: part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int. Nov;107(44):776-82. doi: 10.3238/arztebl.2010.0776. Epub 2010 Nov 5. PMID: 21116397; PMCID: PMC2992018.

[3] Grigorev A, 2020, MLwiki/NormalEquations**,** http://mlwiki.org/index.php/Normal_Equation

[4] Austin D., 2022 Understanding Linear Algebra, Chapter 3.1.4

[5] Draper N. R., Smith H., 1998, Applied Regression Analysis. Third Edition. John Wiley and Sons, Inc, New York, NY, p. 126

[6] Shakhla S, Shah B, Shah N, Unadkat V, Pratik K., 2018. Stock Price Trend Prediction Using Multiple Linear Regression. https://www.ijesi.org/papers/Vol(7)i10/Version-2/D0710022933.pdf