Least Squares Linear Regression Synthesis

Kenza Filali

Mount Royal University

MATH 2303: Linear Algebra for Data Science

Dr. Adam Humeniuk

December 9th, 2025

Contents

0. Introduction

In this project, we will be walking through everything that makes Least Squares Linear Regression one of the most popular and overused models today. The core idea of linear regression is that it finds the best-fitting straight line through a set of data points. You're probably wondering what this infamous line is used for; well, it is used to describe the relationship between two variables, x and y, respectively manipulated and responding. The end goal of this crucial model is to predict the value of one variable based simply on the other from this very line. We then dive into the Coefficient of Determination to see how much variance is impacting results. We will look into the limitations of our regression analysis, also touching on the correlation versus causation ordeal, and wrapping up with the relevance of this model and how prevalent it is in data analysis and in today's world. **I'll be taking an intuitive approach and constantly showing visuals throughout to best understand the following concepts, since math is always best understood when visualized!**

1. Data

I found and chose this student performance dataset from Kaggle: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams . The reason for this choice was due to the high volume of instances(=rows), which would benefit in stronger insights. I was also quite intrigued by reading scores versus writing scores. Growing up, teachers always told us that reading more would aid in our writing. I read a lot but could never quite figure out an essay. Therefore, let us understand everything about regression while evaluating if reading scores can actually aid or predict writing scores. We will be going through the following concepts using this data!

2. Least Squares Regression

As previously mentioned, we want to find the best straight line that falls well between scattered points in order to truly represent a proper meaningful trend. Least squares is a technique for fitting an equation, line, curve, function, or model to a set of data. That said, this has more versatility then the simple linear regression.

IMPORTANT TO NOTE:

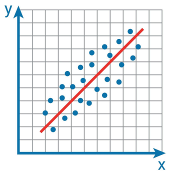
It's worth noting that other equations, such as parabolas and polynomials can also be fit using linear least squares, as long as the variables being optimized are linear.

2.1 Slope line:

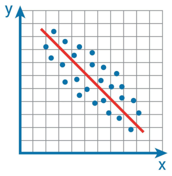
We know that $y = mx + b$ is the equation for a straight line

- m is the slope (how steep the line is)
- b is the y-intercept (where the line crosses the y-axis)

Now, imagine you go out and collect a bunch of real-world data points (like the scores you see plotted below). When you look at those points, they might somewhat look like a line, but they won't fall perfectly on one single line. They'll be scattered around it.



Positive Correlation



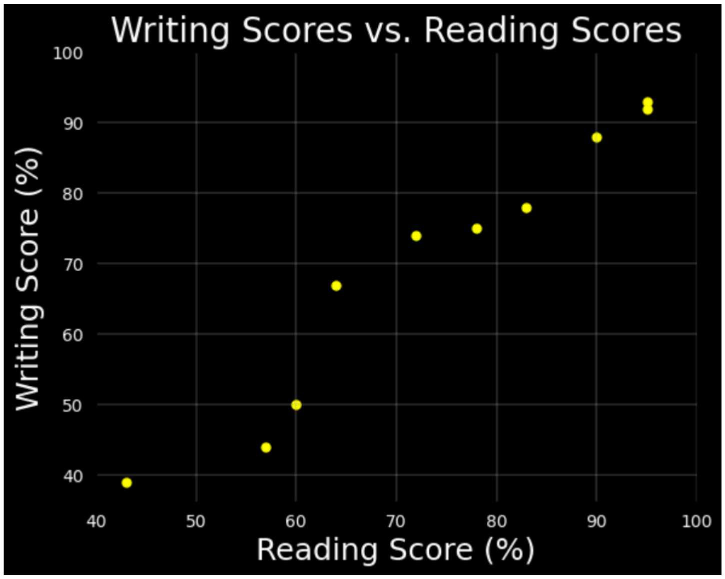
Negative Correlation

The goal of Least Squares is to find the one single line ($y = mx + b$) that comes closest to ALL of those scattered points at the same time. This very line is named the **"line of best fit"** or the **"regression line"**.

2.2 "Least squares concept"

The name "Least Squares" tells you exactly how we define what it means for closest. Let us take a **small sample of our data** to better understand. We will take the **first 10 rows** and plot what that looks like:

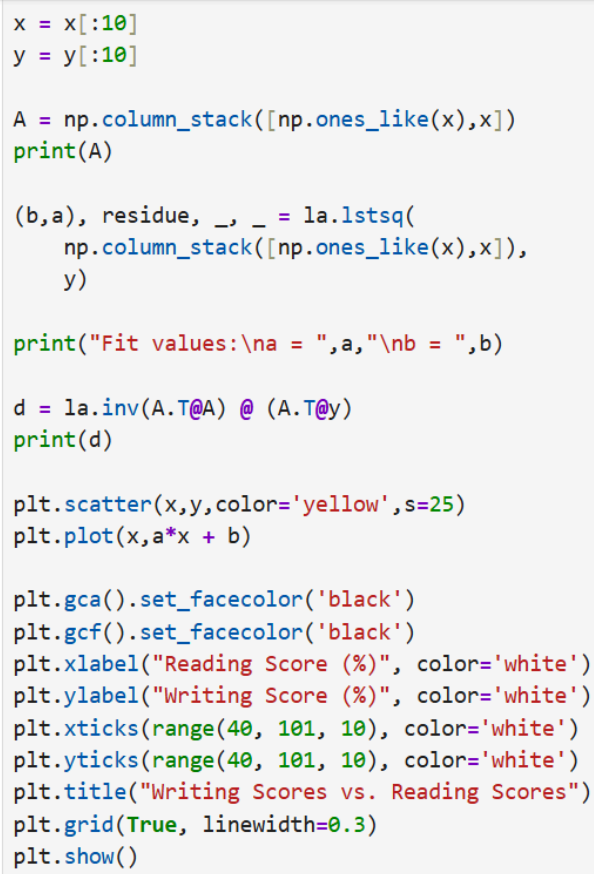
Reading Score(%)	Writing Score(%)
72	74
90	88
95	93
57	44
78	75
83	78
95	92
43	39
64	67
60	50



We can see that all the data points are scattered. Now we want to find the regression line that best fits this data.

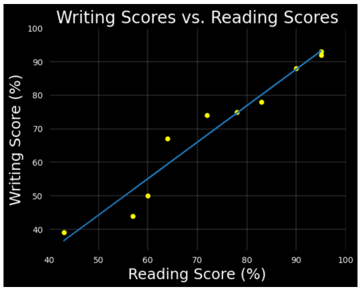
Code Check: 2.2.1 How to find the Least Squares linear fit?

In Numpy:



```
x = x[:10]
y = y[:10]

A = np.column_stack([np.ones_like(x),x])
print(A)

(b,a), residue, _, _ = la.lstsq(
    np.column_stack([np.ones_like(x),x]),
    y)

print("Fit values:\na = ",a,"\nb = ",b)

d = la.inv(A.T@A) @ (A.T@y)
print(d)

plt.scatter(x,y,color='yellow',s=25)
plt.plot(x,a*x + b)

plt.gca().set_facecolor('black')
plt.gcf().set_facecolor('black')
plt.xlabel("Reading Score (%)", color='white')
plt.ylabel("Writing Score (%)", color='white')
plt.xticks(range(40, 101, 10), color='white')
plt.yticks(range(40, 101, 10), color='white')
plt.title("Writing Scores vs. Reading Scores")
plt.grid(True, linewidth=0.3)
plt.show()
```

Output:



Explanation:

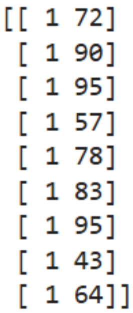
The function la.lstsq() solves the linear least-squares problem:

$$A \bullet x \ \approx y$$

This equation finds the values of x that make the model's predictions as close as possible to the real data.

The line np.column_stack([np.ones_like(x),x]),y)

This creates matrix A:



```
[[ 1 72]
 [ 1 90]
 [ 1 95]
 [ 1 57]
 [ 1 78]
 [ 1 83]
 [ 1 95]
 [ 1 43]
 [ 1 64]]
```

1st column of 1s represents intercept b.

2nd column is variable x, for slope a.

So the model is:
$y = ax + b$

Calling lstsq(A,y):

This solves: $x = (A^T A)^{-1} A^T y$

The tuple of 4 values:

1. (b,a): your best-fit intercept then slope
2. residue: Sum of squared residuals
3. _ : Rank of Matrix A
4. _ : Singular values of A

KNOWLEDGE CHECK-IN
You should now be able to intuitively understand the linear regression line.

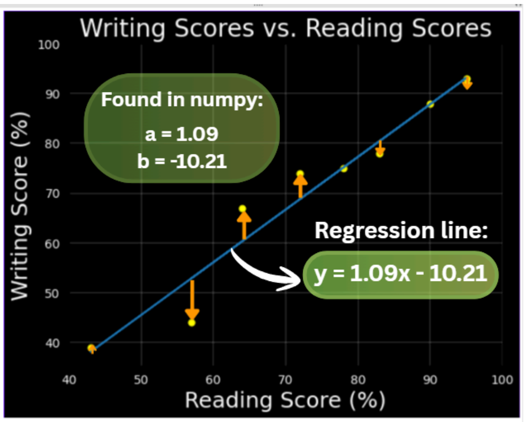
2.3 Everything Error

1. The Error (The "Residual")

Wherever the regression line ends up being, we measure how far the data point is to the line. This vertical line is called the error or the residual.

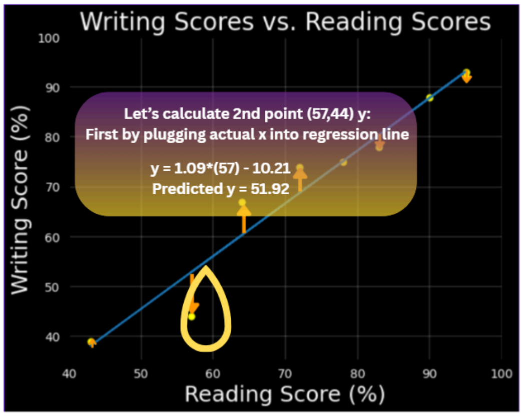
$$Error \ = \ Actual\ y\ value \ - \ Predicted\ y\ value(from\ the\ line)$$

Actual y	Predicted y	Errors
39	36.66	2.34
44	51.92	-7.92
50	55.19	-5.19
67	59.55	7.45
74	68.27	5.73
75	74.81	0.19
78	80.26	-2.26
88	87.89	0.11
92	93.34	-1.34
93	93.34	-0.34



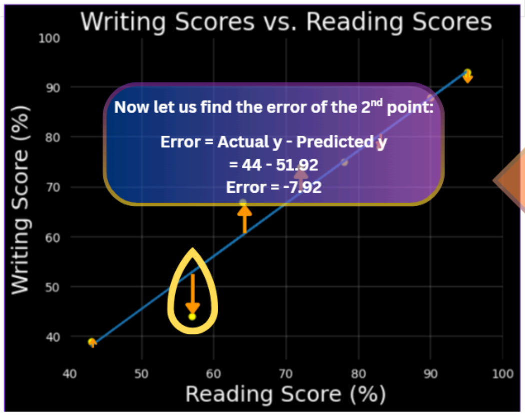
Found in numpy:
a = 1.09
b = -10.21

Regression line:
y = 1.09x - 10.21



Let's calculate 2nd point (57,44) y:
First by plugging actual x into regression line

y = 1.09*(57) - 10.21
Predicted y = 51.92



Now let us find the error of the 2nd point:

Error = Actual y - Predicted y
= 44 - 51.92
Error = -7.92

What if other positive errors cancel the one we just calculated?

WHAT THEN?

2. Squaring the Error(The "Squares")

If a point falls above the line, the error is positive. However, if the point falls below the line it is negative. If we simply added them, the positive and negative errors might cancel each other out. Where does that leave us in regards to accuracy?

Well, to avoid this consequence, we square every error:

$$Squared\ Error \ = \ (Actual\ y \ - \ Predicted\ y)^2$$

- Squaring ensures **all errors become positive**
  - Distance = -5 becomes d = 25.
  - Distance = 5 becomes d = 25.

Errors	2.34	-7.92	-5.19	7.45	5.73	0.19	-2.26	0.11	-1.34	-0.34
Squarred Error	5.48	62.73	26.94	55.50	32.83	0.036	5.11	0.012	1.80	0.116

3. Minimizing the Sum (The "Least")

The final step is to sum up all of these individual squared errors for every single point.

$$Total\ Squared\ Error\ =\ \Sigma((Actual\ y\ -\ Predicted\ y)^2)$$

The **least squares** method is the math technique used to find the values of **m** and **b** that make the total squared error as small as possible. In other words, it finds the line with the "least" error, which gives us a stronger and more reliable result.

By minimizing this total sum, you can guarantee that you have found the absolute best $y\ =\ mx\ +\ b$ line that fits your data.

2.4 Expanding on Matrix Version

We now understand how Least Squares work but now let us see in the lens of Linear Algebra. So let us make sure we truly understand x . We saw in 2.2.1 how to set up the matrix version to solve Least Squares at high-level.

2.4.1 Why do we put all 1s in the first column A?

We know that Ax = b becomes:

$$\begin{array}{cccc} A & x & = & b \\ \begin{vmatrix} 1 & 72 \\ 1 & 90 \end{vmatrix} & \begin{vmatrix} x \\ y \end{vmatrix} & & \begin{vmatrix} 74 \\ 88 \end{vmatrix} \end{array} \equiv \begin{vmatrix} 1x + 72y = 74 \\ 1x + 90y = 88 \end{vmatrix}$$

Does this equation look familiar??

1st Column: Therefore we emulate the regression line formula by making the x a constant by always having the factor 1 which will represent our *b*.

2nd Column: As we can see in the expanded form, the second column becomes the factor to our m or a.

We can see that by solving for $b(= x)$ and $a(= y)$, we can find a potential relationship equation that relates to our original data for the manipulated variable("Reading Score") and see if we can closely predict/output the responding variable("Writing Score").

2.4.2 Now solving for $x = (A^T A)^{-1} A^T y$:

We know the basis is $Ax = b$. Although since this original equation has no exact solution, we then multiply both sides by A^T. Now that we can solve it directly, this new equation will always have a solution. Fortunately, that solution is the one that makes the error as small as possible.

$$\boxed{A^T Ax = A^T b}$$

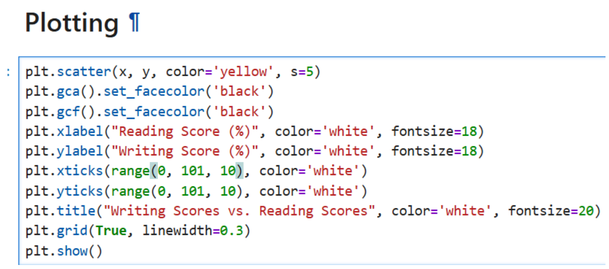
This new equation is very favorable to use, having a square matrix($A^T A$) is beneficial being both symmetric and invertible. We then solve for vector x which will give our b and a. We divide the square matrix(or invert) on both sides:

$$\boxed{x = (A^T A)^{-1} A^T\ b}$$

To summarize, least squares is a method often used for fitting a model to data. Residuals express the error between the current model fit and the data. The objective of least squares is to minimize the sum of the squared error across all the data points to find the best fit for a given model.

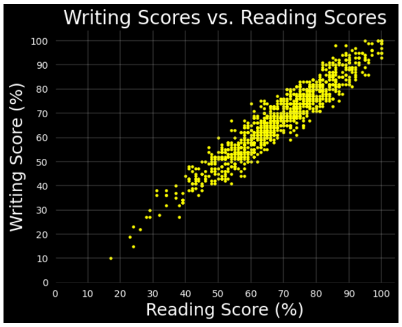
2.5 Application to Real Data:

Importing ALL the data and plotting a scatter plot to visualize all our data points for Writing Scores versus Reading Scores:

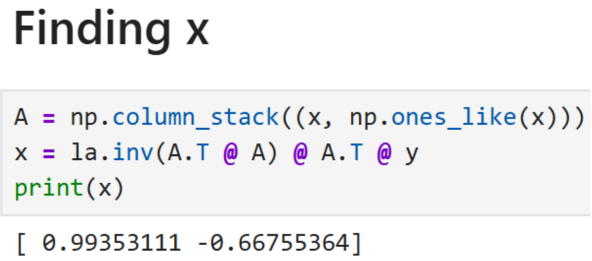


Plotting ¶

```
plt.scatter(x, y, color='yellow', s=5)
plt.gca().set_facecolor('black')
plt.gcf().set_facecolor('black')
plt.xlabel("Reading Score (%)", color='white', fontsize=18)
plt.ylabel("Writing Score (%)", color='white', fontsize=18)
plt.xticks(range(0, 101, 10), color='white')
plt.yticks(range(0, 101, 10), color='white')
plt.title("Writing Scores vs. Reading Scores", color='white', fontsize=20)
plt.grid(True, linewidth=0.3)
plt.show()
```



Remembering to solve for x(= $(A^T A)^{-1} A^T\ b$), we first have to set up for A.



Finding x

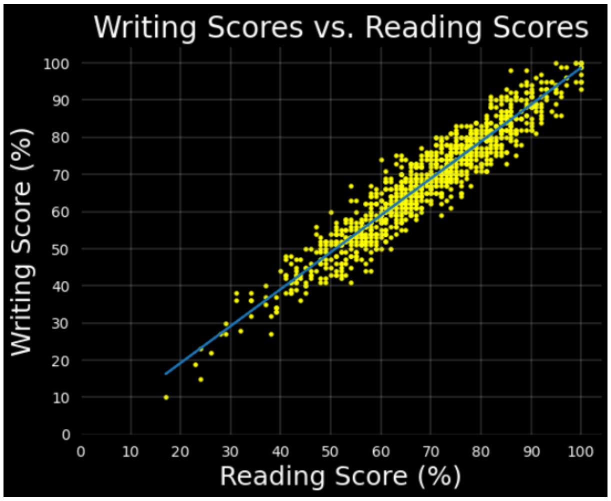
```
A = np.column_stack((x, np.ones_like(x)))
x = la.inv(A.T @ A) @ A.T @ y
print(x)
```

```
[ 0.99353111 -0.66755364]
```

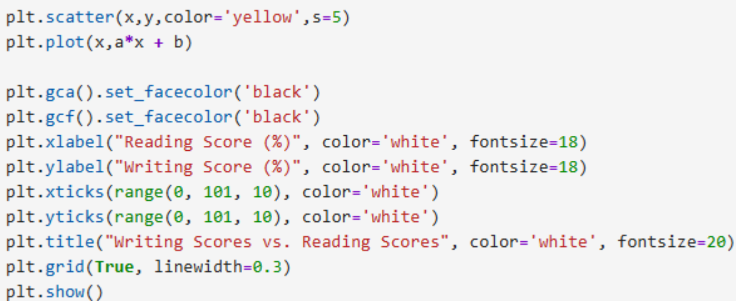
We solve for x=[m,b]^T

Therefore the solution is:

$$\hat{y} = 0.99x - 0.67$$



Plotting our new line within the original data to visualize the solution and regression line



```
plt.scatter(x,y,color='yellow',s=5)
plt.plot(x,a*x + b)

plt.gca().set_facecolor('black')
plt.gcf().set_facecolor('black')
plt.xlabel("Reading Score (%)", color='white', fontsize=18)
plt.ylabel("Writing Score (%)", color='white', fontsize=18)
plt.xticks(range(0, 101, 10), color='white')
plt.yticks(range(0, 101, 10), color='white')
plt.title("Writing Scores vs. Reading Scores", color='white', fontsize=20)
plt.grid(True, linewidth=0.3)
plt.show()
```

3. Correlation of Determination

Once we find the regression line using Least Squares, we want to know if that line is actually decent at explaining your data. That's where the **Coefficient of Determination(** r^2**)** comes in.

From your Intro to Statistics course, we already know about correlation (r).

- We know that the value can only be between -1 and 1. And if it is very close to either bounds, it tells us they are strongly related. Although the closer that r is to 0, the weaker our correlation. Refer back to 2.1 for more information.

We can also bring all the sum of squares from earlier(2.3.3) and use the average value of y. By dividing these two and subtracting from 1, this also gives us r^2.

$$r^2 = 1 - \frac{Sum\ of\ Squared\ Error}{Sum\ of\ average\ y}$$

How Strong is Our Line?

Now squaring r gives us the Coefficient of Determination(r^2), a value between 0 and 1. This factor focuses on measuring the impact and demonstrates how predictable our responding variable is. Now since squaring any number gives us a positive one, our indicators slightly change. For r^2, if it is closer to 0 it is a weak relationship and if it is closer to 1, it is a strong relationship.

This number tells us the proportion of the variation in the dependent variable (y) that can be predicted from the independent variable (x). So in Layman's terms, it tells us how well x can predict y.

💡 Think of it as a percentage score for your line of best fit.

3.1 Connection to Data

r^2 for our data is 0.95.

This means that 95% of the differences we see in students' writing scores can be predicted or explained by looking at their reading scores alone. The remaining 5% of the variation is caused by other factors we didn't include, such as study habits, teaching quality, personal interest, or even random variation. A high r^2 shows that reading scores are a very strong predictor of writing scores in this dataset.

4. Limitations

Limitations in our analysis are crucial to acknowledge even if our results lead to a strong outcome like a high r^2.

1. Generalization:

We can only analyze the data set that was provided. Our insights are only derived from this specific group of students. We essentially cannot assume this exact relationship holds true for other students in different countries, grade levels or all students.

2. Outliers:

The data points that are far away from the main cluster are called outliers. These outliers can affect the Least Squares line and the line loses accuracy. Therefore, outliers could misrepresent the relationship for the majority of students based on our given data.

3. True Scope:

Our very last plausible limitation is that correlation does not imply causation. More on this in the next section.

Overall, understanding these limitations helps us stay honest about what our results really mean. Even if our r^2 is high, it only reflects this specific group of students. By keeping in mind things like limited data, outliers, and the fact that correlation doesn't prove causation, we avoid making claims that are too big or misleading. This ensures our conclusions stay accurate and responsible.

5. Correlation vs Causation

The most significant limitation is that our analysis only shows a correlation, a relationship between Reading Scores and Writing Scores. However, this does not entirely prove that high reading ability *causes* high writing ability.

It’s important to remember that just because two things happen at the same time doesn’t mean one caused the other. There’s a famous example of ice-cream sales and drowning deaths increasing together. It might look like ice-cream somehow leads to drowning, but really a third factor is hot weather, which makes people both buy more ice-cream and go swimming.

If people confuse correlation with causation, it can lead to wrong ideas, bad decisions, and potential fear mongering. It can also make the public stop trusting research when those mistakes are discovered. This is why understanding that correlation does not mean causation is essential for good research and for avoiding harmful misunderstandings.

6. World Relevance

Least squares linear regression is widely used because it helps us find clear, reliable patterns in messy real-world data. It gives us a straightforward way to model relationships and make informed decisions. By minimizing error and providing the line of best fit, it allows scientists, businesses, and governments to analyze data accurately and avoid guessing.

Today, with AI and huge amounts of data everywhere, this method matters more than ever. It’s one of the basic tools that data analysts and AI systems use to learn from information, spot trends, and make smart decisions in areas like business, health, finance, and technology. Without simple, reliable methods like least squares, modern data analysis and many AI models wouldn’t work nearly as well.

7. References

1. Quantfish publisher. (2022). *Linear Regression.* [Video recording]. Quantfish.
2. (2025). Youtube.com. https://www.youtube.com/watch?v=S0ptaAXNxBU
3. Cui, C., & Fearn, T. (2017). Comparison of partial least squares regression, least squares support vector machines, and Gaussian process regression for a near infrared calibration. *Journal of near Infrared Spectroscopy (United Kingdom)*, *25*(1), 5–14. https://doi.org/10.1177/0967033516678515
4. *Least Squares Regression*. (2023). Mathsisfun.com. https://www.mathsisfun.com/data/least-squares-regression.html
5. *The Method of Least Squares*. (2025). Gatech.edu. https://textbooks.math.gatech.edu/ila/least-squares.html
6. Quantfish publisher. (2021). *Linear Regression Analysis in Mplus.* [Video recording]. Quantfish.
7. to, C. (2002, September 8). *approximation method in statistics*. Wikipedia.org; Wikimedia Foundation, Inc. https://en.wikipedia.org/wiki/Least_squares
8. Zelin, A. (2017). *Call center forecasting : linear regression models.* [Video recording]. SAGE Publications Ltd.
9. (2025). Youtube.com.https://www.youtube.com/watch?v=8B271L3NtAw
10. (2025). Youtube.com.https://www.youtube.com/watch?v=GtV-VYdNt_g
11. *Ordinary Least Squares regression (OLS)*. (2025). XLSTAT, Your Data Analysis Solution. https://www.xlstat.com/solutions/features/ordinary-least-squares-regression-ols
12. Quantfish publisher. (2022). *Demonstration & Practice: Fitting Linear Regression Models with R.* [Video recording]. Quantfish.

13. *Khan Academy*. (2023). Khanacademy.org. https://www.khanacademy.org/math/ap-statistics/bivariate-data-ap/xfb5d8e68:residuals/v/regression-residual-intro

14. GeeksforGeeks. (2023, July 6). *Least Square Method | Definition Graph and Formula*. GeeksforGeeks. https://www.geeksforgeeks.org/maths/least-square-method/