MULTIVARIABLE LINEAR REGRESSION MODEL

Aidan M. Andrucyk, aa1918@scarletmail.rutgers.edu

Purpose: This personal project affords the opportunity to synthesize prior academic knowledge from disciplines including linear algebra, statistics, and computer science while also applying newfound knowledge gained from further investigation. If you have any comments, suggestions, or questions, please feel free to email me at aa1918@scarletmail.rutgers.edu.

Abstract: The program uses least squares approximation to generate a linear regression model with response variable $\hat{y}$ and multiple user-specified predictor variables.

Background: This background guide assumes a foundational understanding of linear algebra on the readers part, but attempts to offer extensive explanations of each component. This background guide will also be preferential towards single explanatory variable conditions due to their simplicity in implementation, intuitive geometric interpretation, and overall ease of understanding for the reader.

Foundational concepts must be understood prior to delineating specifics of the program:

Calculus

Partial Derivative: the derivative of a multivariable expression is its derivative with respect to a single variable, holding all other variables constant.

Linear Algebra

Vector: an array data structure consisting of elements of like kind.

Dot Product: the sum of the products of the corresponding entries with respect to two vectors.

Matrix: a rectangular array (two-dimensional grid) arranged in rows and columns consisting of elements of like kind.

Identity Matrix: a square matrix with values of one along the diagonal and values of zero along the off-diagonal.

Transpose: an operator which flips a matrix along the diagonal, switching the row and column entries of a given matrix.

Inverse: the corresponding square matrix which produces the identity matrix through matrix multiplication with the original matrix with a nonzero determinant.

Determinant: a scalar value that describes the linear transformation of a matrix

Cofactor: the determinant of a square submatrix which is derived from removing the respective column and row.

Cofactor Expansion (Laplace Expansion): a method for deriving the determinant of a matrix using cofactors.

Adjugate: the transpose of the cofactor matrix.

Subspace: a space that is a subset of some larger space that contains the zero vector and is closed under both addition and multiplication.

Statistics

Explanatory (Predictor) Variable: a type of independent variable that can still be affected by other factors.

Responsive Variable: a type of dependent variable that is affected by explanatory variables.

Error: the deviation of the observed value from the *true* value.

Residual: the difference between the observed value and the *estimated* value.

Variance: the distance of a set of numbers from their average value.

Covariance: the joint variability of two variables.

Variance–Covariance Matrix: a matrix with variances in the diagonal and off-diagonals have the covariance of each element with itself.

Determination coefficient: the proportion of the responsive variable's variance that can be predicted give the explanatory variable(s).

Correlation coefficient: square root of that coefficient of determination that expresses the relationship between two variables. A high correlation coefficient does not necessarily imply that there is a casual relationship.

Extrapolation: an estimation outside the range of the original data set based on the data within the range.

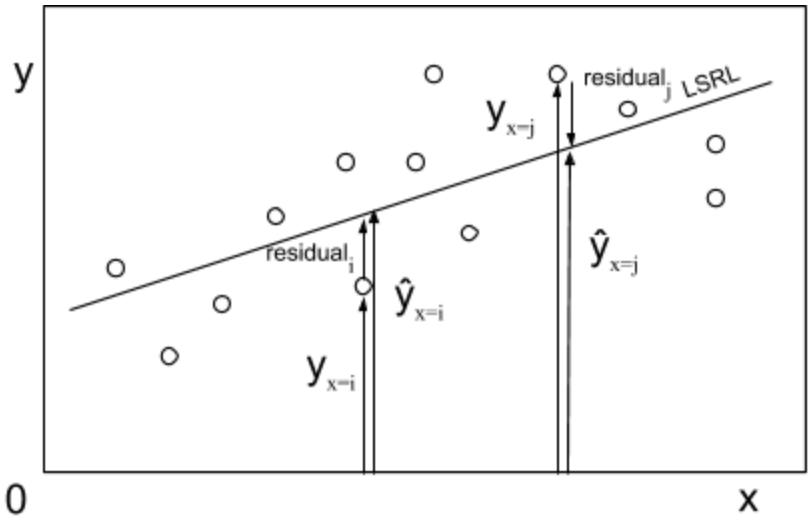
Regression Analysis: a statistical process for estimating the relationships between a response variable and predictor variable(s).

Linear Regression Model: a linear regression analysis for a response variable and predictor variable(s).

Least Squares Approximation: method of estimating the mathematical relationship between the predicting and responsive variables that minimizes the summation of the squared residuals. In the linear regression framework, this relationship is expressed as a Least Squares Regression Line (LSRL) or, colloquially, a line of best fit. The LSRL general formula is as follows $\hat{y} = b_0 + b_1x_1 + \dots + b_kx_k$, where k equals the number of predictor variables. The following is a simple, yet instructive conceptual derivation of the least squares approximation with a single explanatory variable:

Data Set: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

Imperfect Geometric Interpretation:



Symbolic Derivation:

Design Matrix $C = \{v_1, v_k\}$, where $v_1 = \{1_1, 1_2, \dots, 1_n\}$ and $v_2 = \{x_1, x_2, \dots, x_n\}$, where C is invertible;

Response Vector $\dot{y} = \{y_0, y_1, \dots, y_n\}$;

Parameter Vector $\beta = \{b_0, b_1\}$;

Projection Matrix $P_w = C(C^TC)^{-1}C^T$;

Residual $= y_{x=i} - \hat{y}_{x=i}$, where $\hat{y}_{x=i} = b_0 + b_1x_n$ and i/j correspond to their respective data point $\therefore$

Residual Vector $= \{y_1 - \hat{y}_1, y_2 - \hat{y}_2, \dots, y_n - \hat{y}_n\}$ $\therefore$

Minimum of the sum of the squared residuals $= \Sigma_1^n \|\text{distance from } \dot{y} \text{ to the span}\{b_0, b_1\}\|^2$, where the span$\{b_0, b_1\}$ is a subspace $= \Sigma_1^n(\text{orthogonal projection of } \dot{y} \text{ to the span}\{b_0, b_1\})^2 = \Sigma_1^n(y_{x=k} - \hat{y}_{x=k})^2 = \Sigma_1^n(y_{x=k} - b_0 - b_1x_n)^2 = b_0v_0 + b_1v_1 = C\beta$ $\therefore$

$C\beta = P_w\dot{y}$ $\therefore$

$C\beta = C(C^TC)^{-1}C^T\dot{y}$ $\therefore$

$C^TC\beta = C^TC(C^TC)^{-1}C^T\mathring{y}$, where $C^TC(C^TC)^{-1} = I_n$ ∴

$C^TC\beta = I_nC^T\mathring{y}$ ∴

$\beta = (C^TC)^{-1}C^T\mathring{y}$ ∴

The least squares approximation can be derived using the form $\beta = (C^TC)^{-1}C^T\mathring{y}$ such that $\hat{y} = b_0 + b_1x_1$.

However, this program expands upon data data which is contingent upon a single explanatory variable, and allows for several explanatory variables, which is to say extends from simple linear regression to multiple linear regression. The form of multiple linear regression is intuitively $\hat{y} = b_0 + b_1x_1 + \ldots + b_kx_k$, where k equals the number of predictor variables and can be further inferred from a simple dimensional analysis of the form $\beta_{k+1*1} = (C_{k+1*n}{}^TC_{n*k+1})^{-1}C_{k+1*n}{}^T\mathring{y}_{n*1}$, where the end product will have matching vectors of form $(k+1)_x1$.

Since $\hat{y} = b_0 + b_1x_1 + \ldots + b_kx_k$ represents a linear function with the first term of degree equal to 0 and one for all terms thereafter, the partial derivative with respect to each of the explanatory variables tells us that the corresponding parameter is the differential coefficient, or the difference in the response variable for every marginal change in the respective predictor. It is clear that b_0 is geometrically the y-intercept, or the predicted value of y when all of the explanatory variables are equal to zero.

Limitations of least squares approximation include but not limited to the sensitivity to outliers, the contingency upon the invertibility of $(C^TC)^{-1}$, and the assumption linear relationship exists between the responding and predicting variables with negligible correlation among the predicting variables.

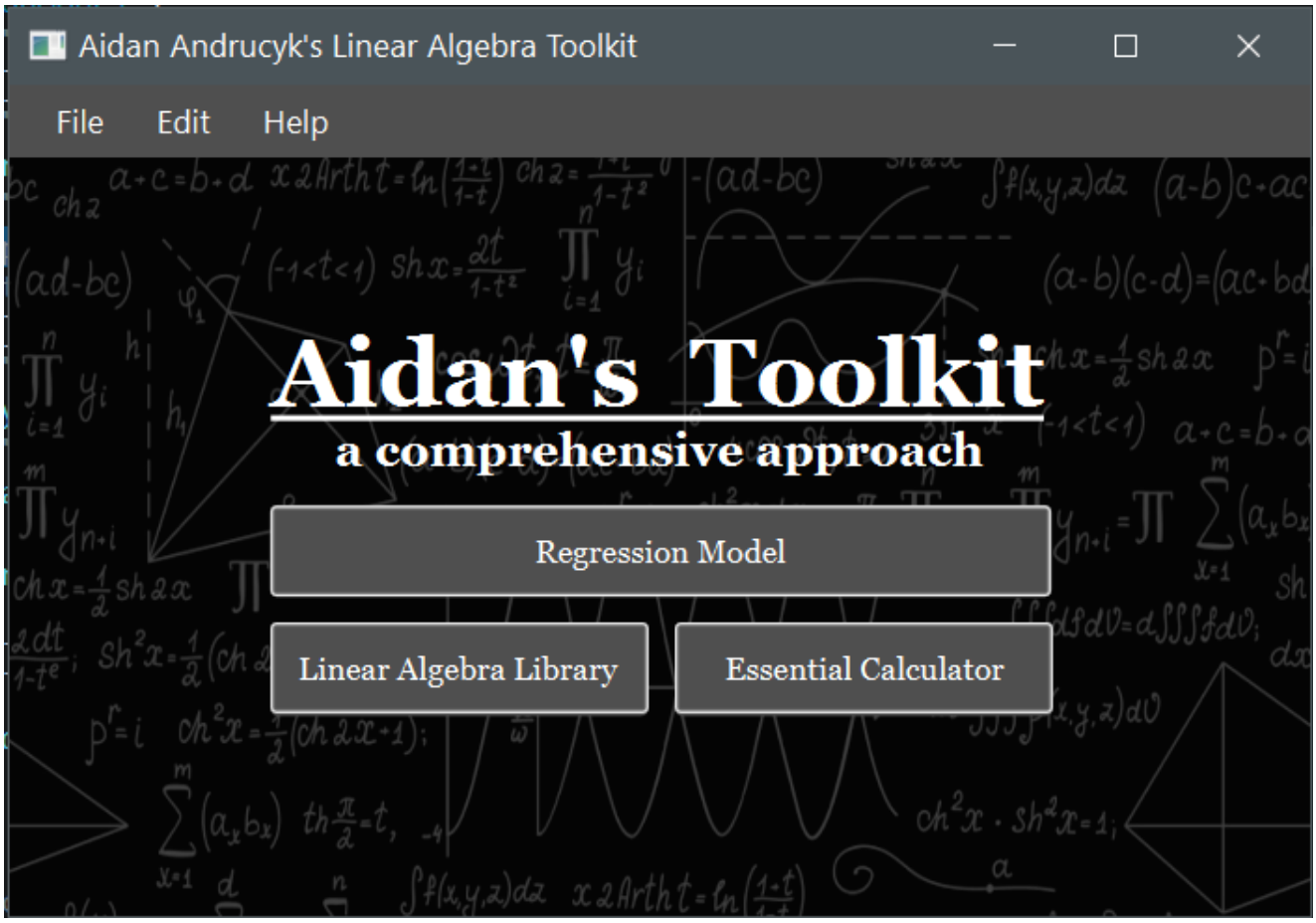
Practical Applications of Multiple Linear Regression:

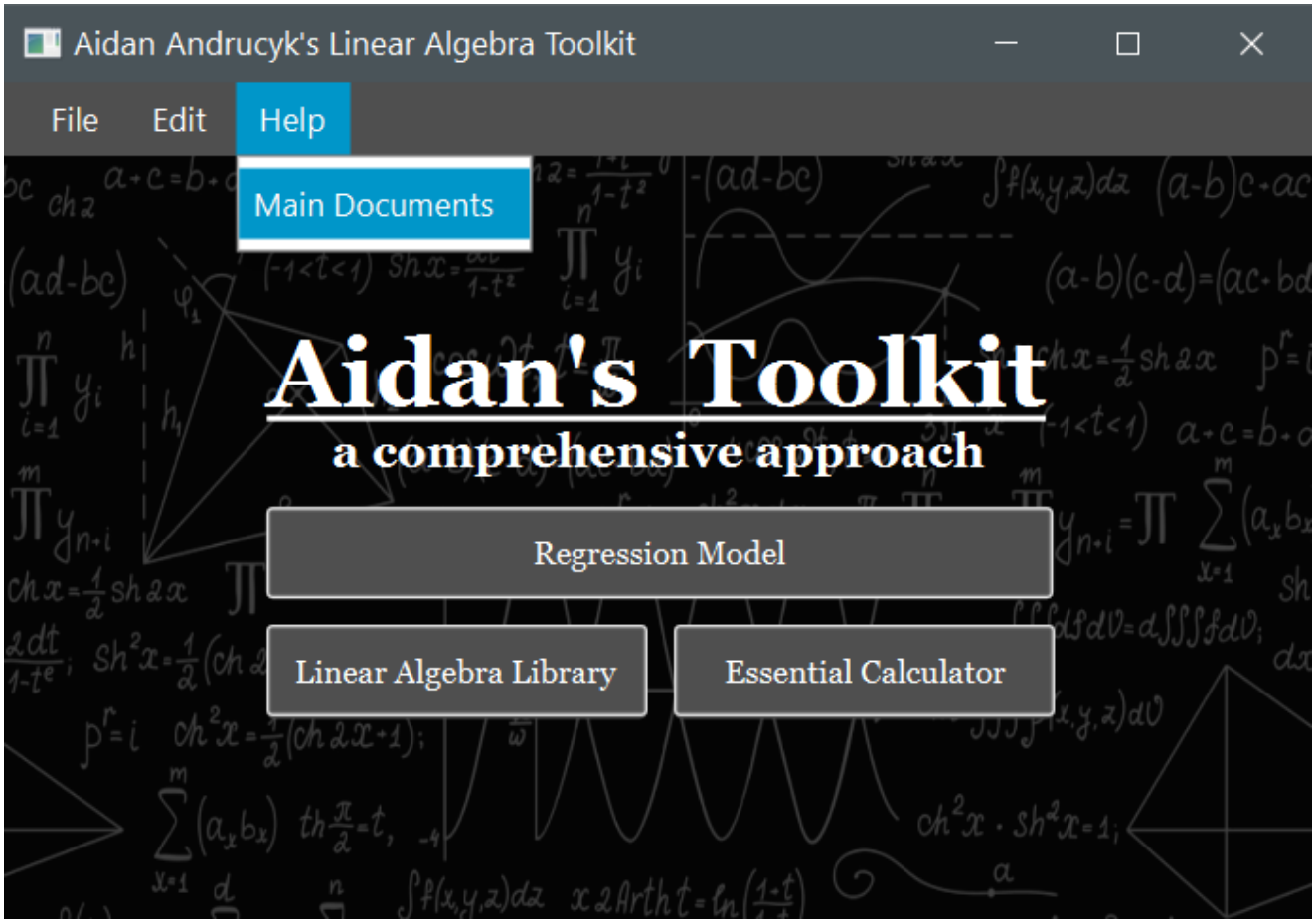
I. Predicting the price of ExxonMobil Stock prices with predictors of interest rates, oil prices, value of S&P 500 index, price of oil futures, etc.

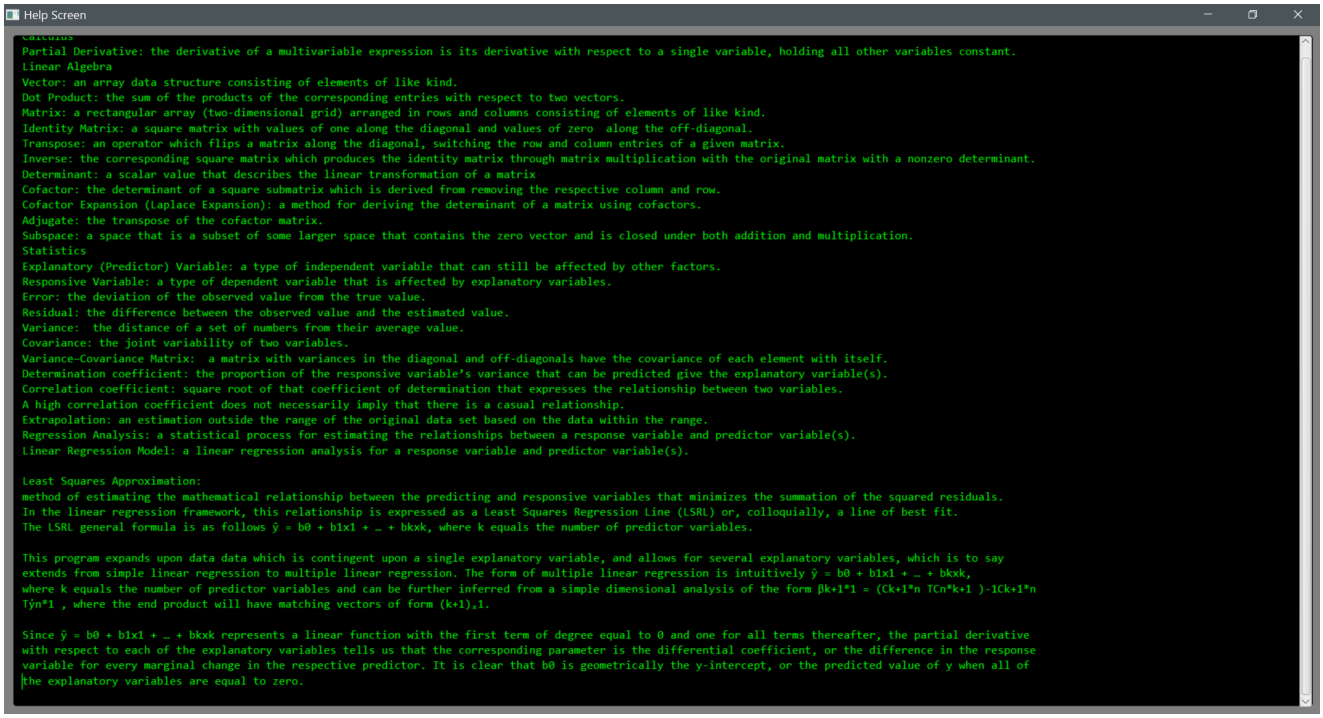
II. Extrapolating the average yearly temperature in a given region with predictor(s) CO2, average windy intensity, cloud cover rates, etc.

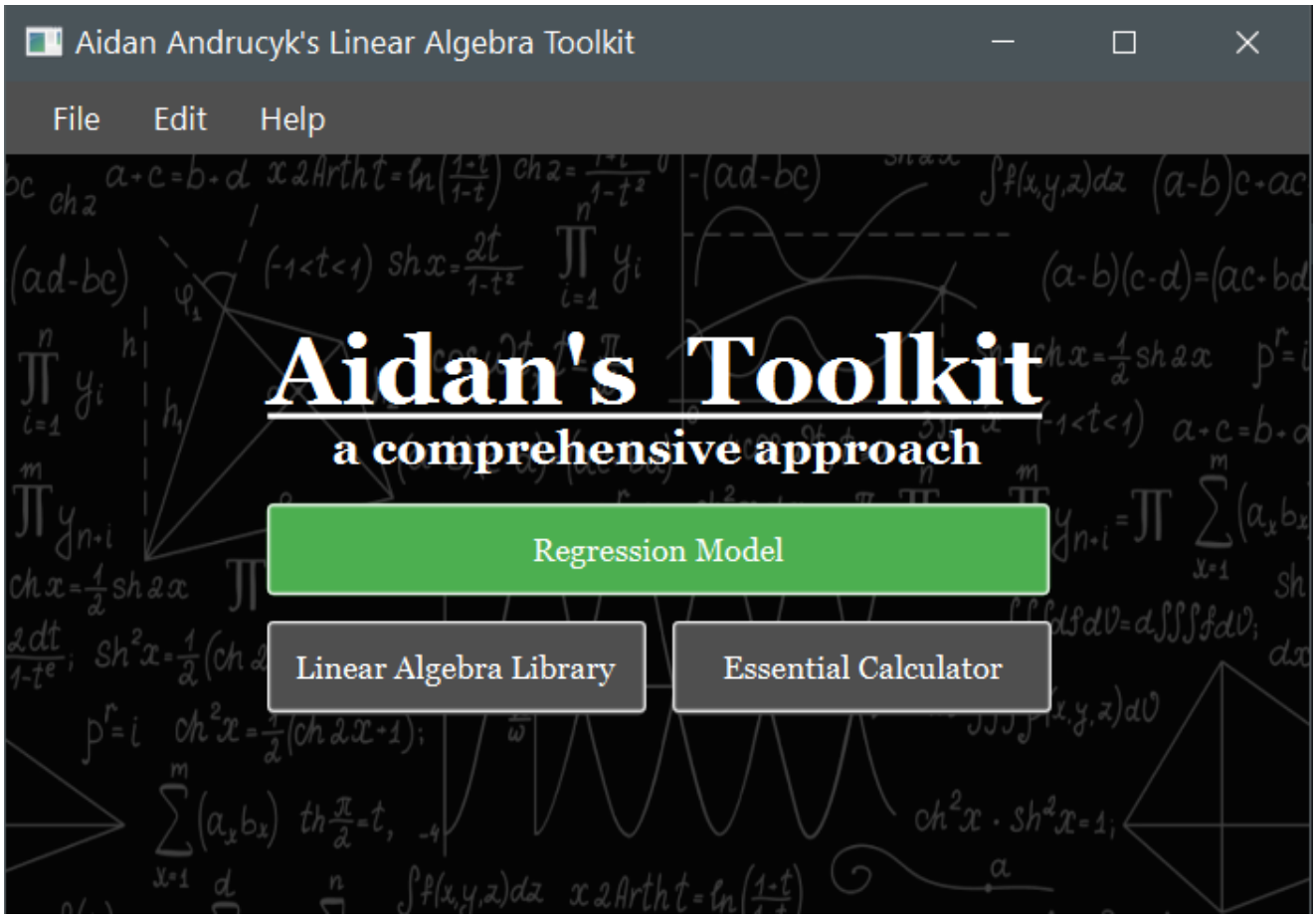
III. Estimating the price of a home in a given location with predictor(s) of construction year, lot size, number of rooms, etc.

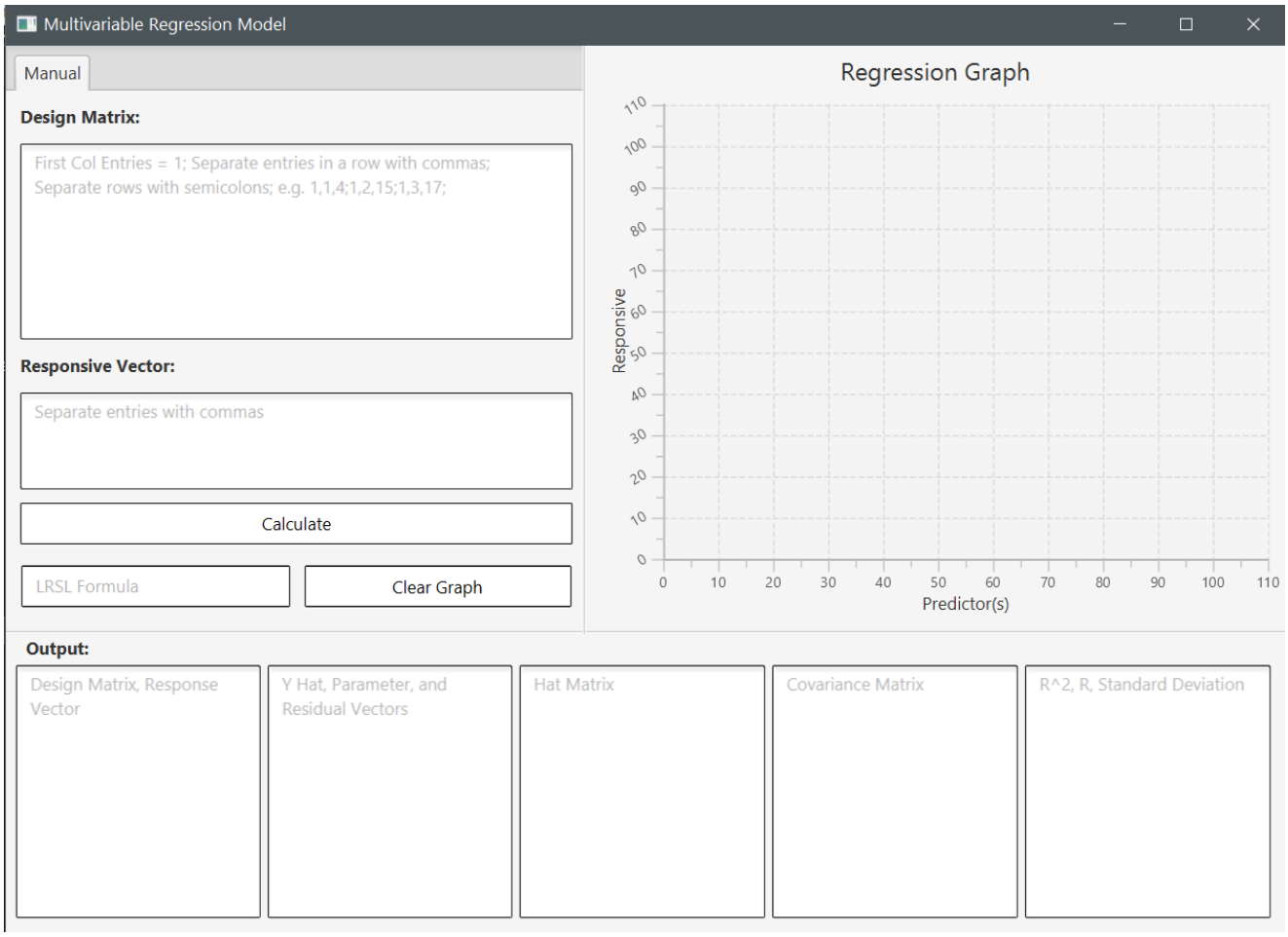
Demonstration:



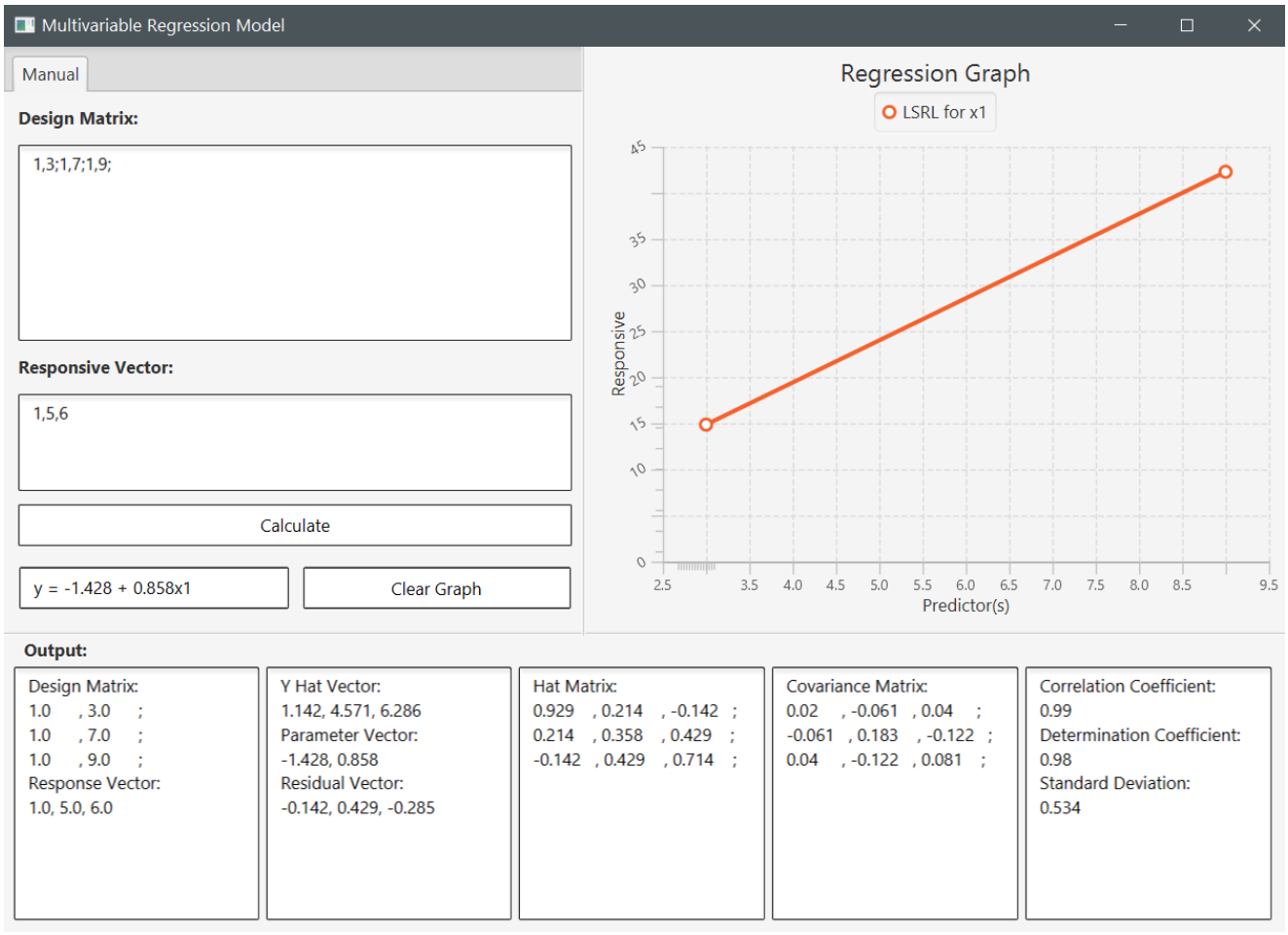

Aidan Andrucyk's Linear Algebra Toolkit
File
Edit
Help
Main Documents
Aidan's Toolkit
a comprehensive approach
Regression Model
Linear Algebra Library
Essential Calculator

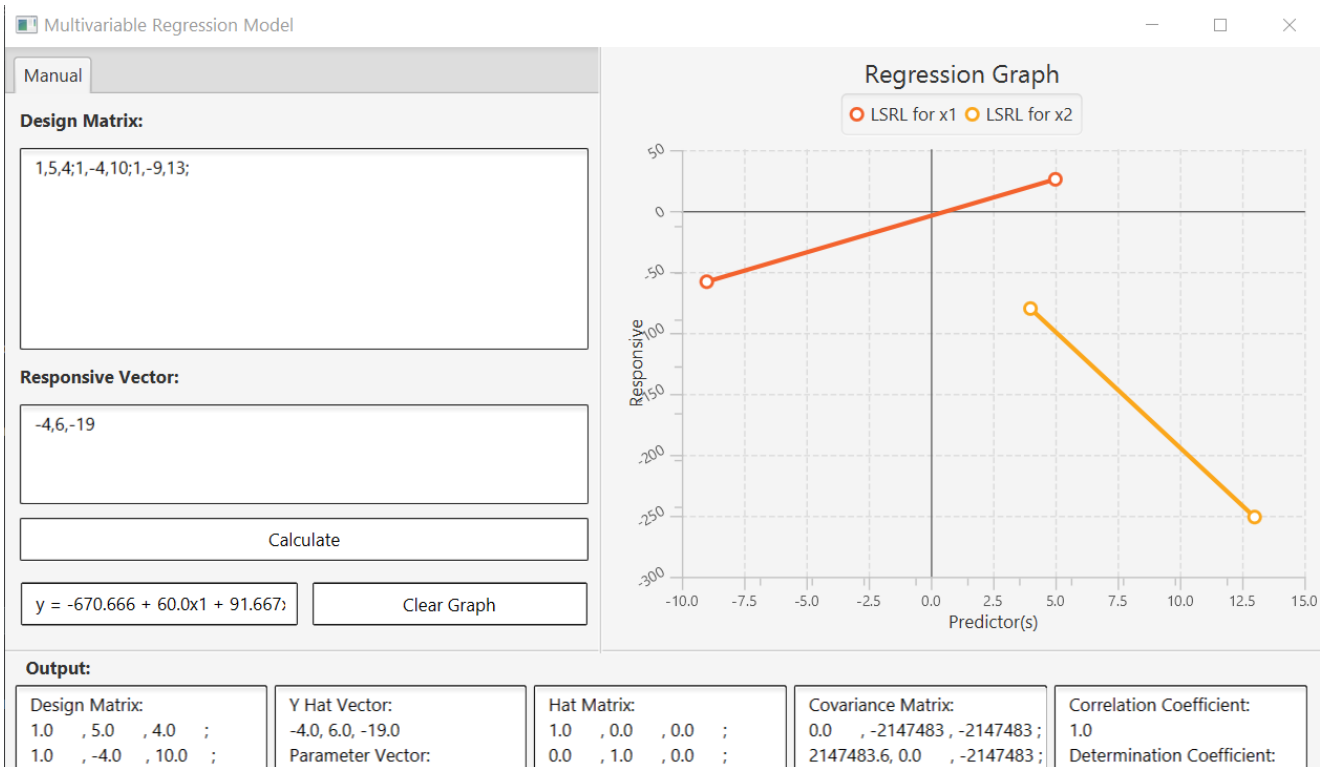

Help Screen
Partial Derivative: the derivative of a multivariable expression is its derivative with respect to a single variable, holding all other variables constant.
Linear Algebra
Vector: an array data structure consisting of elements of like kind.
Dot Product: the sum of the products of the corresponding entries with respect to two vectors.
Matrix: a rectangular array (two-dimensional grid) arranged in rows and columns consisting of elements of like kind.
Identity Matrix: a square matrix with values of one along the diagonal and values of zero along the off-diagonal.
Transpose: an operator which flips a matrix along the diagonal, switching the row and column entries of a given matrix.
Inverse: the corresponding square matrix which produces the identity matrix through matrix multiplication with the original matrix with a nonzero determinant.
Determinant: a scalar value that describes the linear transformation of a matrix
Cofactor: the determinant of a square submatrix which is derived from removing the respective column and row.
Cofactor Expansion (Laplace Expansion): a method for deriving the determinant of a matrix using cofactors.
Adjugate: the transpose of the cofactor matrix.
Subspace: a space that is a subset of some larger space that contains the zero vector and is closed under both addition and multiplication.
Statistics
Explanatory (Predictor) Variable: a type of independent variable that can still be affected by other factors.
Responsive Variable: a type of dependent variable that is affected by explanatory variables.
Error: the deviation of the observed value from the true value.
Residual: the difference between the observed value and the estimated value.
Variance: the distance of a set of numbers from their average value.
Covariance: the joint variability of two variables.
Variance-Covariance Matrix: a matrix with variances in the diagonal and off-diagonals have the covariance of each element with itself.
Determination coefficient: the proportion of the responsive variable's variance that can be predicted give the explanatory variable(s).
Correlation coefficient: square root of that coefficient of determination that expresses the relationship between two variables.
A high correlation coefficient does not necessarily imply that there is a casual relationship.
Extrapolation: an estimation outside the range of the original data set based on the data within the range.
Regression Analysis: a statistical process for estimating the relationships between a response variable and predictor variable(s).
Linear Regression Model: a linear regression analysis for a response variable and predictor variable(s).
Least Squares Approximation:
method of estimating the mathematical relationship between the predicting and responsive variables that minimizes the summation of the squared residuals.
In the linear regression framework, this relationship is expressed as a Least Squares Regression Line (LSRL) or, colloquially, a line of best fit.
The LSRL general formula is as follows ŷ = b0 + b1x1 + … + bkxk, where k equals the number of predictor variables.
This program expands upon data data which is contingent upon a single explanatory variable, and allows for several explanatory variables, which is to say
extends from simple linear regression to multiple linear regression. The form of multiple linear regression is intuitively ŷ = b0 + b1x1 + … + bkxk,
where k equals the number of predictor variables and can be further inferred from a simple dimensional analysis of the form βk+1*1 = (Ck+1*n TCn*k+1)-1Ck+1*n
Tŷn*1 , where the end product will have matching vectors of form (k+1),1.
Since ŷ = b0 + b1x1 + … + bkxk represents a linear function with the first term of degree equal to 0 and one for all terms thereafter, the partial derivative
with respect to each of the explanatory variables tells us that the corresponding parameter is the differential coefficient, or the difference in the response
variable for every marginal change in the respective predictor. It is clear that b0 is geometrically the y-intercept, or the predicted value of y when all of
the explanatory variables are equal to zero.

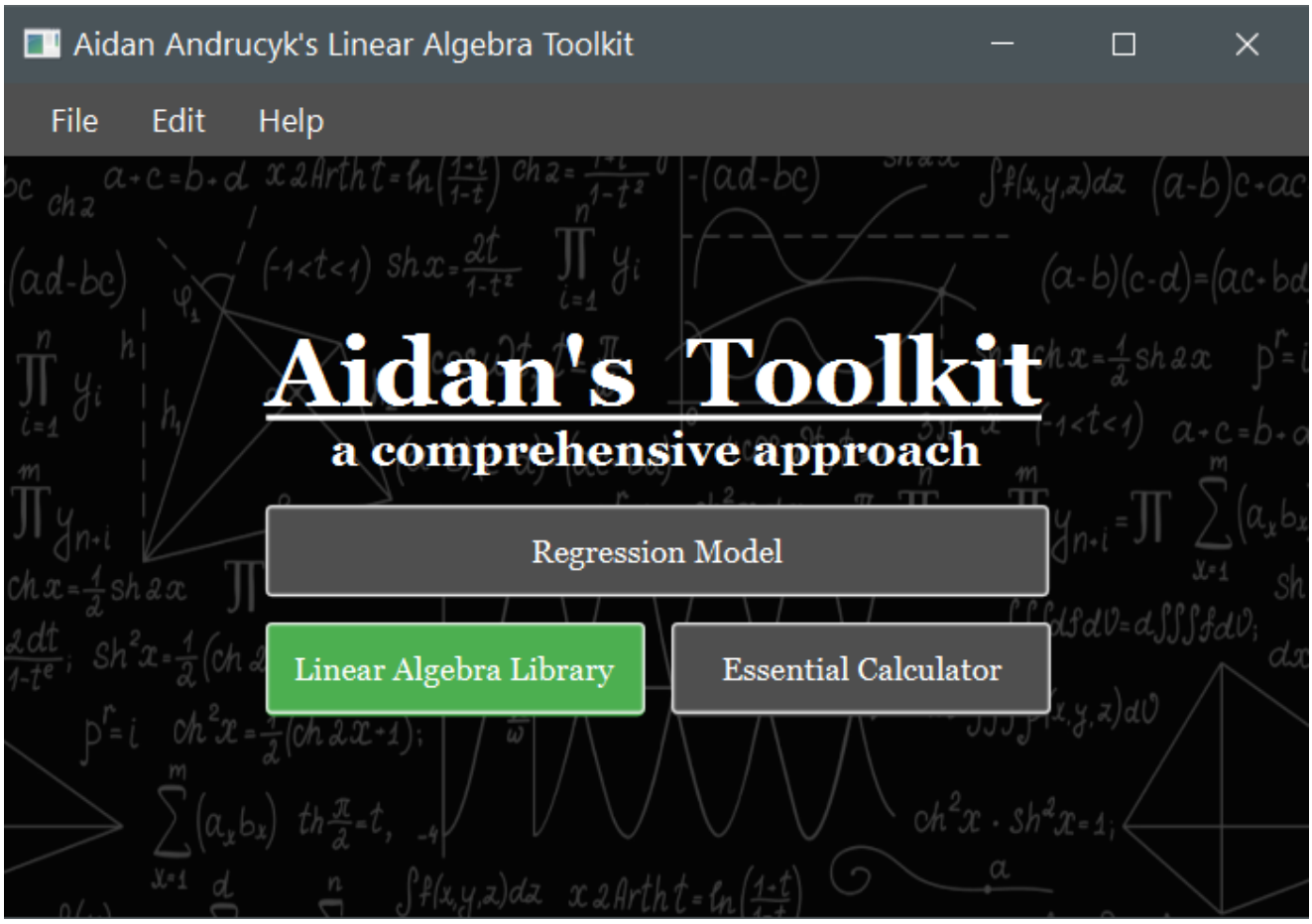

Aidan Andrucyk's Linear Algebra Toolkit
File
Edit
Help
Aidan's Toolkit
a comprehensive approach
Regression Model
Linear Algebra Library
Essential Calculator

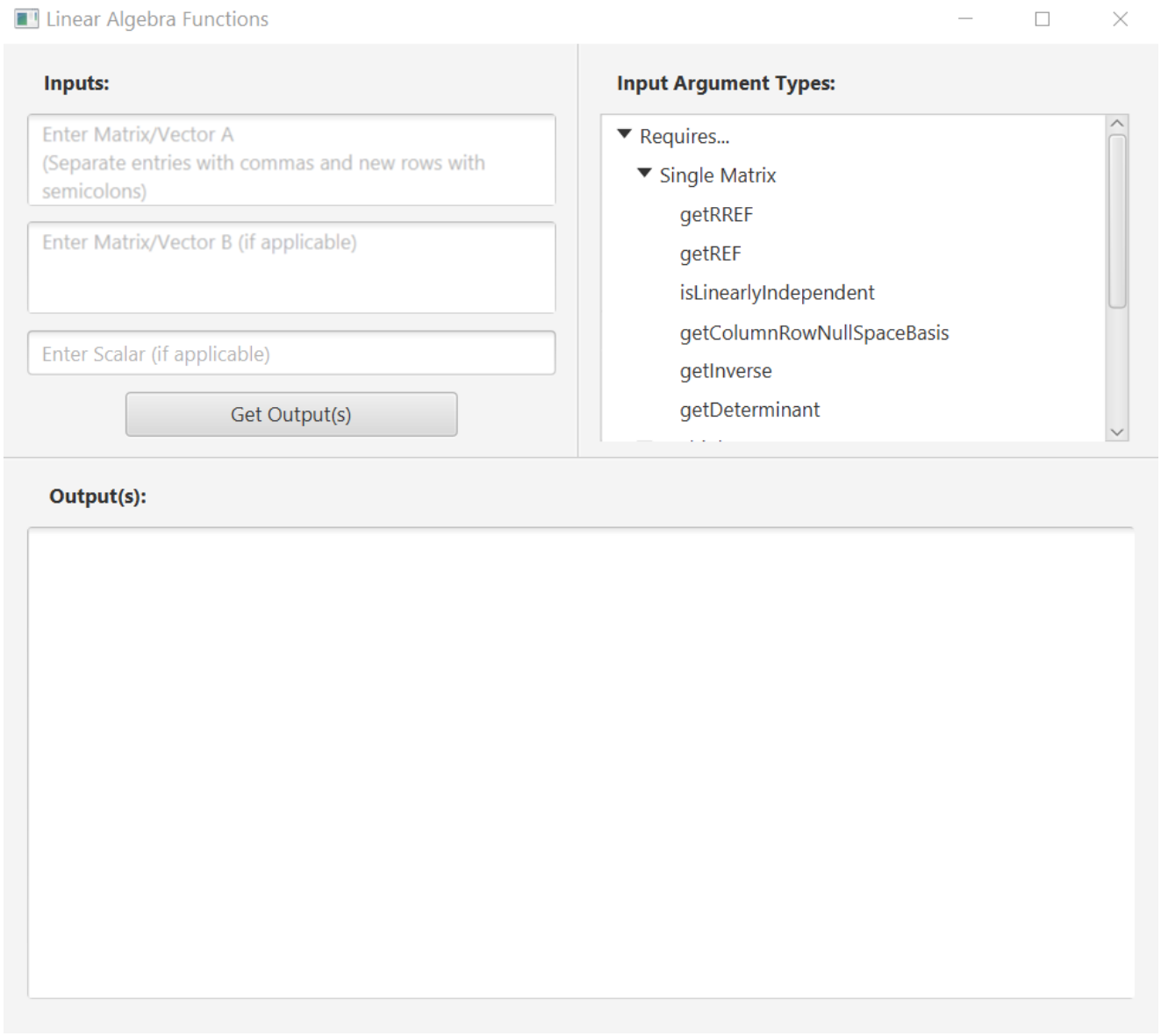

Multivariable Regression Model
Manual
Design Matrix:
First Col Entries = 1; Separate entries in a row with commas;
Separate rows with semicolons; e.g. 1,1,4;1,2,15;1,3,17;
Responsive Vector:
Separate entries with commas
Calculate
LRSL Formula
Clear Graph
Regression Graph
Responsive
Predictor(s)
0
10
20
30
40
50
60
70
80
90
100
110
Output:
Design Matrix, Response Vector
Y Hat, Parameter, and Residual Vectors
Hat Matrix
Covariance Matrix
R^2, R, Standard Deviation

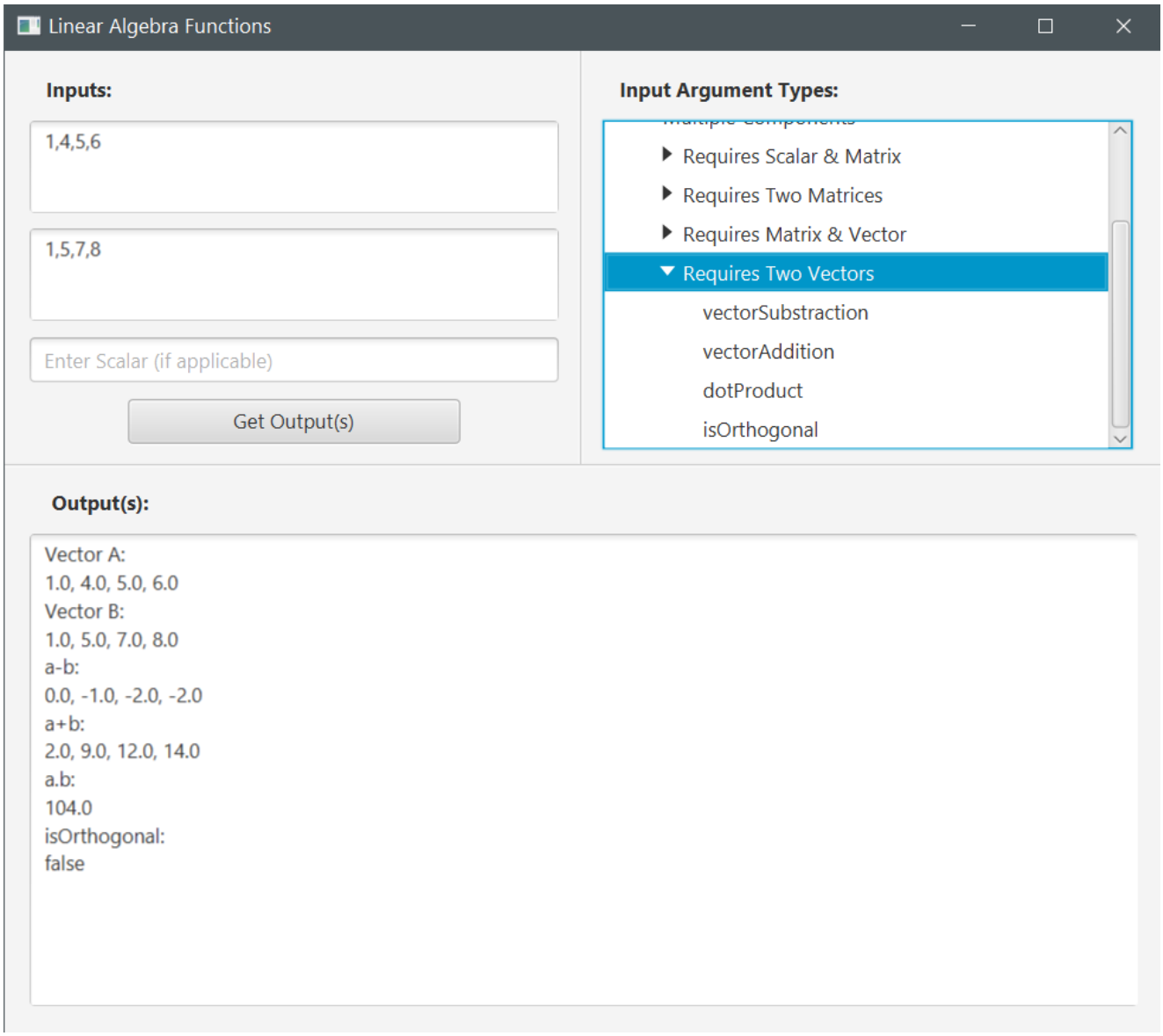
Notice that the LSRL only graphs within the range of the observations

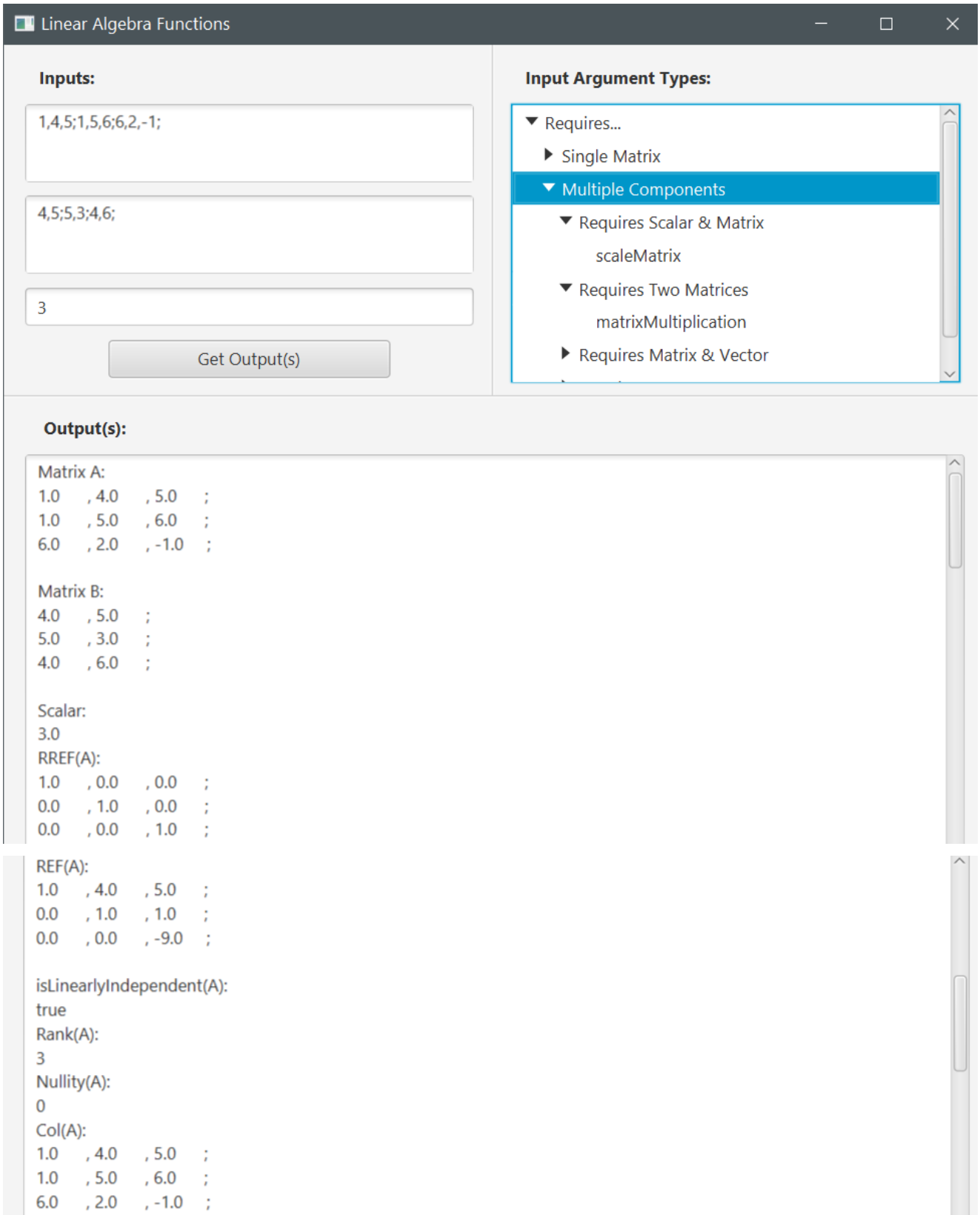





Aidan Andrucyk's Linear Algebra Toolkit
File
Edit
Help
Aidan's Toolkit
a comprehensive approach
Regression Model
Linear Algebra Library
Essential Calculator

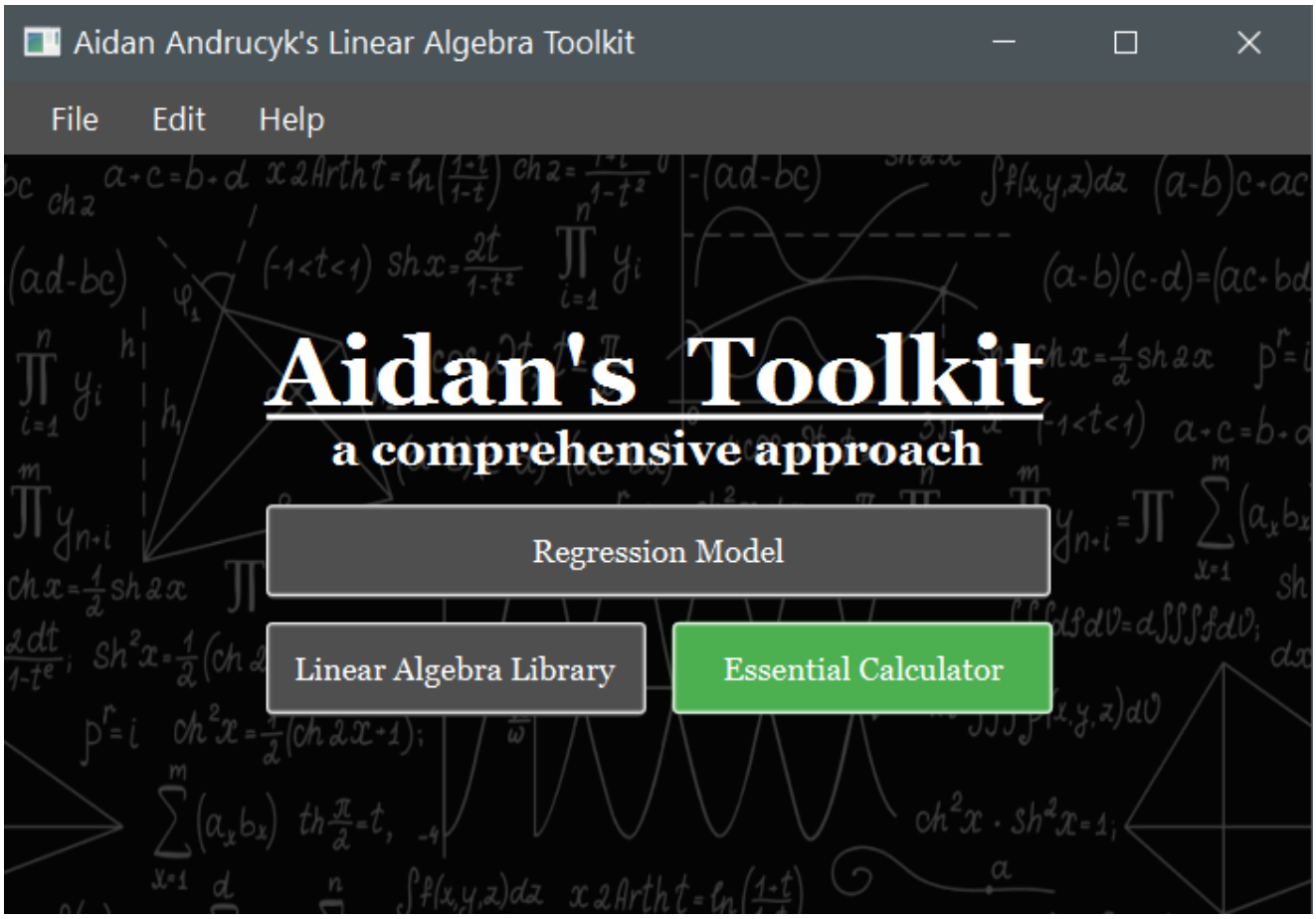

Linear Algebra Functions
Inputs:
Enter Matrix/Vector A
(Separate entries with commas and new rows with semicolons)
Enter Matrix/Vector B (if applicable)
Enter Scalar (if applicable)
Get Output(s)
Input Argument Types:
Requires...
Single Matrix
getRREF
getREF
isLinearlyIndependent
getColumnRowNullSpaceBasis
getInverse
getDeterminant
Output(s):

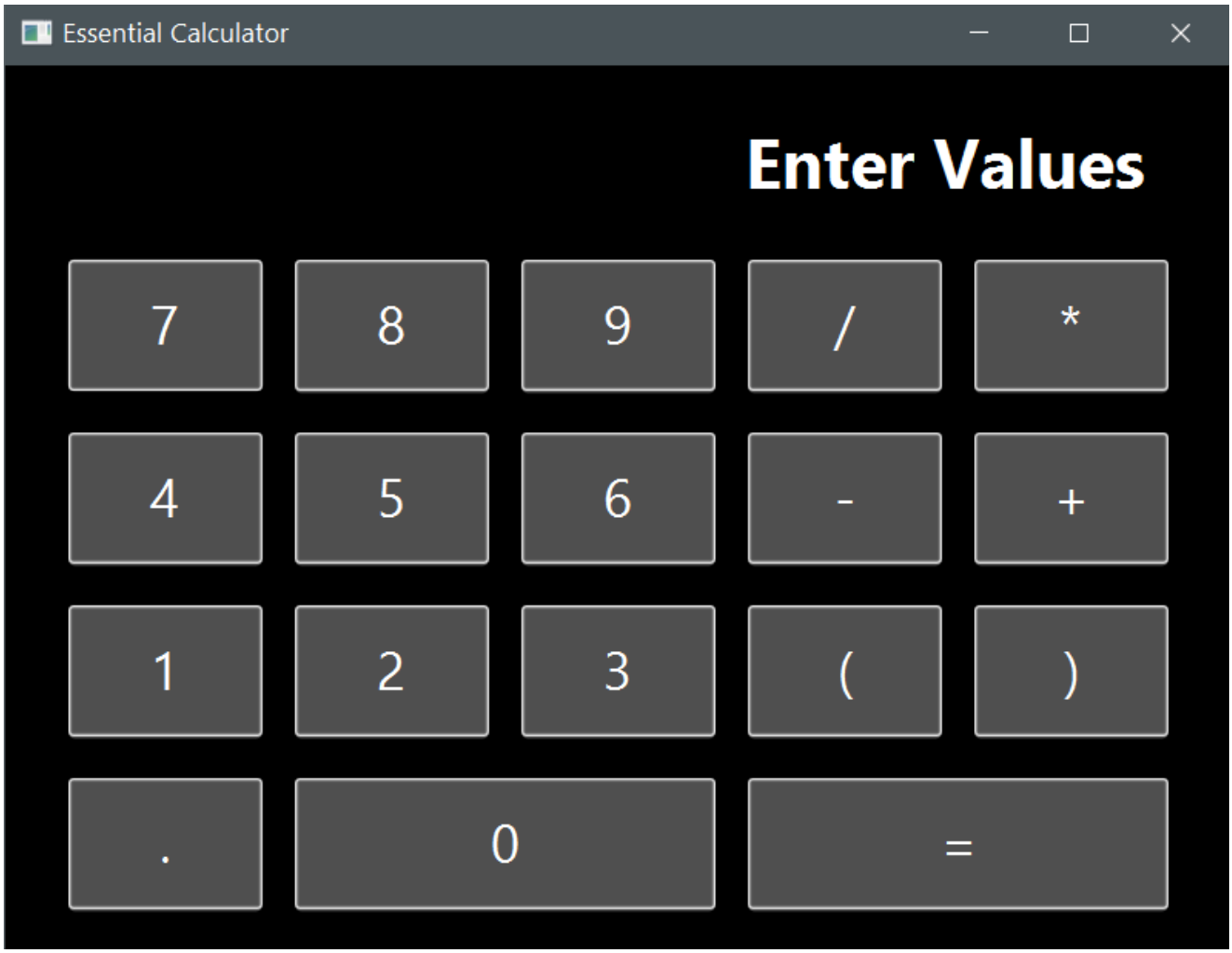

Linear Algebra Functions
Inputs:
1,4,5,6
1,5,7,8
Enter Scalar (if applicable)
Get Output(s)
Input Argument Types:
Requires Scalar & Matrix
Requires Two Matrices
Requires Matrix & Vector
Requires Two Vectors
vectorSubstraction
vectorAddition
dotProduct
isOrthogonal
Output(s):
Vector A:
1.0, 4.0, 5.0, 6.0
Vector B:
1.0, 5.0, 7.0, 8.0
a-b:
0.0, -1.0, -2.0, -2.0
a+b:
2.0, 9.0, 12.0, 14.0
a.b:
104.0
isOrthogonal:
false

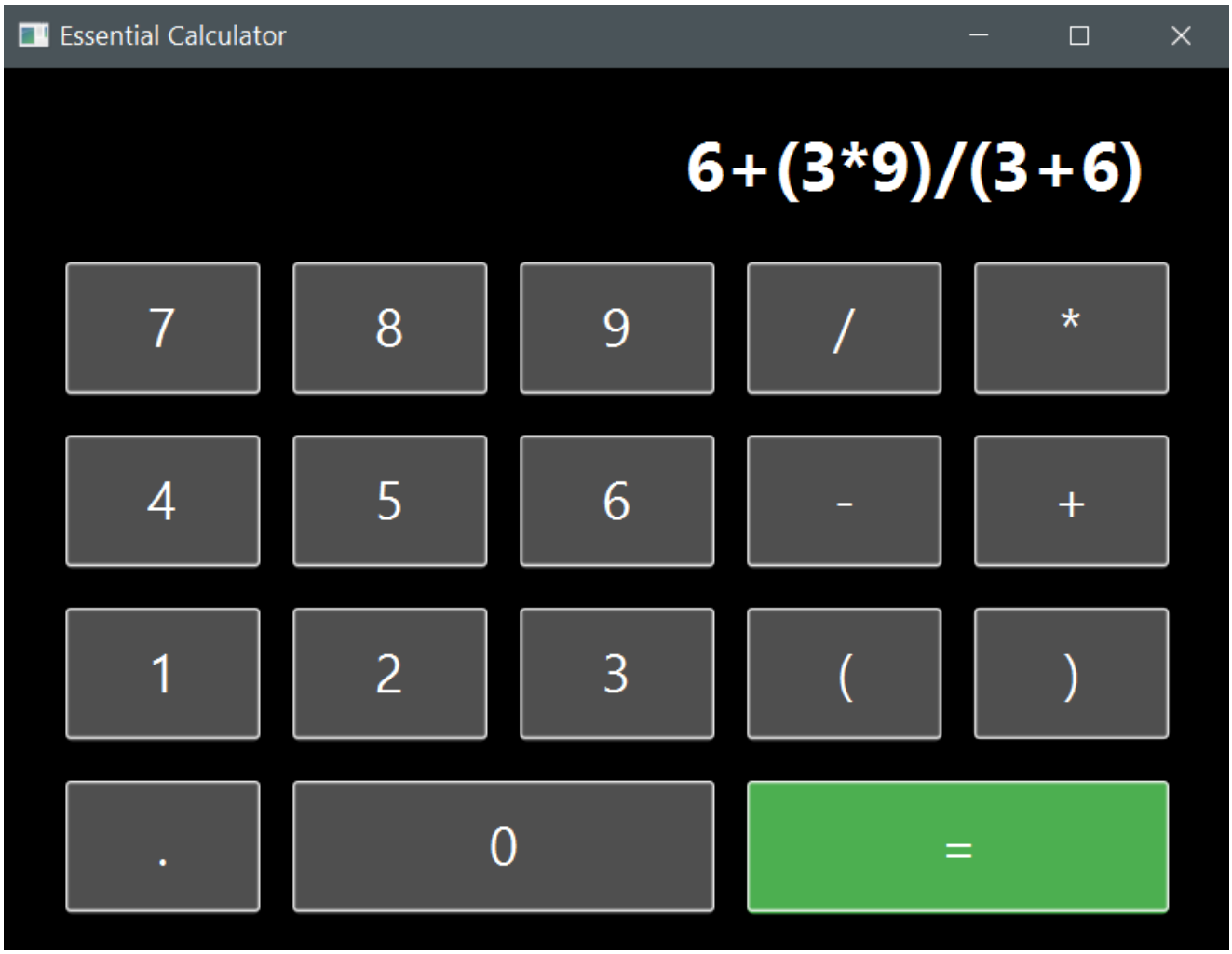

Linear Algebra Functions
Inputs:
1,4,5;1,5,6;6,2,-1;
4,5;5,3;4,6;
3
Get Output(s)
Input Argument Types:
Requires...
Single Matrix
Multiple Components
Requires Scalar & Matrix
scaleMatrix
Requires Two Matrices
matrixMultiplication
Requires Matrix & Vector
Output(s):
Matrix A:
1.0 , 4.0 , 5.0 ;
1.0 , 5.0 , 6.0 ;
6.0 , 2.0 , -1.0 ;
Matrix B:
4.0 , 5.0 ;
5.0 , 3.0 ;
4.0 , 6.0 ;
Scalar:
3.0
RREF(A):
1.0 , 0.0 , 0.0 ;
0.0 , 1.0 , 0.0 ;
0.0 , 0.0 , 1.0 ;
REF(A):
1.0 , 4.0 , 5.0 ;
0.0 , 1.0 , 1.0 ;
0.0 , 0.0 , -9.0 ;
isLinearlyIndependent(A):
true
Rank(A):
3
Nullity(A):
0
Col(A):
1.0 , 4.0 , 5.0 ;
1.0 , 5.0 , 6.0 ;
6.0 , 2.0 , -1.0 ;

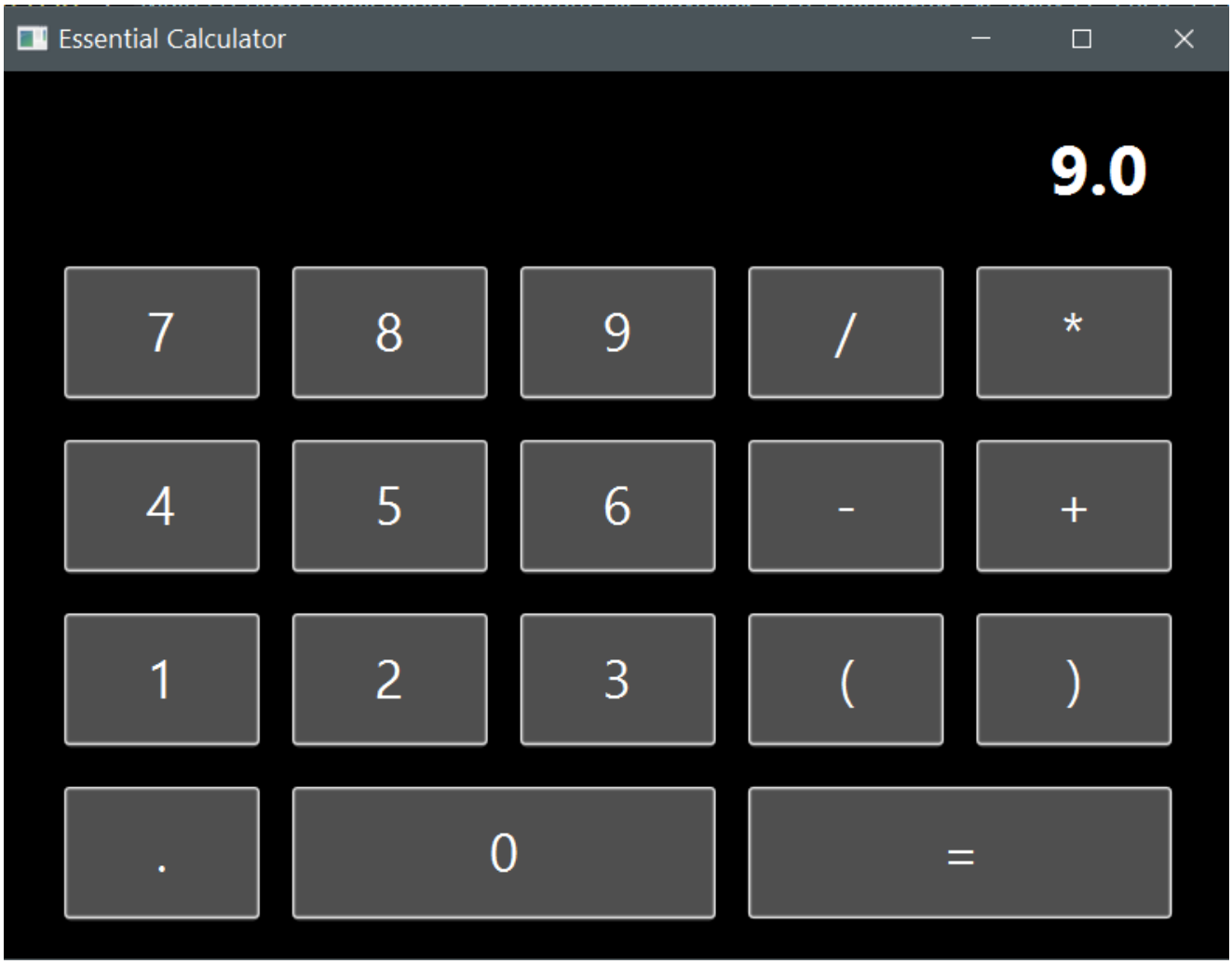
```
Row(A):
1.0     , 1.0     , 6.0     ;
4.0     , 5.0     , 2.0     ;
5.0     , 6.0     , -1.0    ;

Transpose(A):
1.0     , 1.0     , 6.0     ;
4.0     , 5.0     , 2.0     ;
5.0     , 6.0     , -1.0    ;

Determinant(A):
-9.0
Inverse(A):
1.889    , -4.111   , 3.111    ;
-1.555   , 3.444    , -2.444   ;
0.111    , 0.111    , -0.111   ;
```

```
Scaled(A):
3.0     , 12.0    , 15.0    ;
3.0     , 15.0    , 18.0    ;
18.0    , 6.0     , -3.0    ;

AB:
132.0   , 141.0   ;
159.0   , 168.0   ;
90.0    , 90.0    ;
```


Aidan Andrucyk's Linear Algebra Toolkit
File
Edit
Help
Aidan's Toolkit
a comprehensive approach
Regression Model
Linear Algebra Library
Essential Calculator


Essential Calculator
Enter Values
7
8
9
/
*
4
5
6
-
+
1
2
3
(
)
.
0
=


Essential Calculator
6+(3*9)/(3+6)
7
8
9
/
*
4
5
6
-
+
1
2
3
(
)
.
0
=


Essential Calculator
9.0
7
8
9
/
*
4
5
6
-
+
1
2
3
(
)
.
0
=

Technical Components:

I. Language(s): Java, CSS, FXML
II. Library/Libraries: JavaFX
III. Software: Eclipse IDE, Gluon's SceneBuilder

Stylistic Choices:

I. Rounds to the nearest thousandths place for the end of all computations.
    A. Excessively long values might hinder readability and adversely affect the user experience, especially if a single LRSL equation is spanning several lines. The thousandths place provides relatively precise data without sacrificing a large degree of insight.
    B. The use of the double data type to hold values leads to imperfect precision in certain operations, particularly with irrational numbers. For instance, the fraction 4/21 is different from a double storing the value corresponding to 4 divided by 21 which is 0.19047619047619047. The error derives from the double data type's lack of storage availability for the full value. Although this may seem negligible for a single operation, several operations like this compounding on each other may lead to some imprecision which is often the case with least square approximations. With a simple data set such as [(1,14), (3, 17), (5, 19), (7, 20)], an exact LSRL would be $\hat{y} = 13.5 + 1.0x_1$ whereas an algorithm storing imprecise double values outputs $\hat{y} = 13.499999999999998 + 1.0x_1$. Rounding to the thousandths solves specific cases such as the aforementioned data set.
    C. Rounding at the end of processes prevents premature rounding from distorting data integrity.
    D. One unimplemented solution could be the fractional storage of values in an object with two attributes: a numerator double and a denominator double. There are a number of libraries that afford this capability. However, the marginal cost of this implementation far exceeds the additional marginal insight added.
II. Creating a certain static "make[insert class attribute variable]" methods in the RegressionModel class rather than deriving all attributes within the constructor so as to enhance readability with clearer connections to in-line annotations.
III. Ordering method arguments in descending memory allocation order.
IV. Distinguishing vectors as one-dimensional arrays and matrices as two-dimensional arrays. Although possible to allocate vectors as a two-dimensional array of row/column length one, vectors are instead allocated with double[] for memory but largely aesthetic and personal cognition reasons.
V. Avoiding the use of the term "sigma" to avoid confusion.
    A. Denoting the square root of mean squared error (MSE) as "Standard Deviation" rather than lowercase sigma .
    B. Denoting running sums as "sums of..." rather than "sigma of…" or Σ.
VI. Multi-line comments, capitalized letter first word, and periods for annotations as long or longer than three lines; end-line comments, uncapitalized letters, and no periods for annotations shorter than three lines.

Acknowledgements:

1. Lasantha Goonetilleke, Assistant Professor of the Mathematics Department at Rutgers University
2. Bari Leff, AP Statistics Teacher at East Brunswick High School
3. Jose Gomez, JavaFX Instructor

Works Consulted:

Findsen, L. (n.d.). Statistics 512: Applied Linear Models. Retrieved May 16, 2020, from https://www.stat.purdue.edu/~boli/stat512/lectures/topic3.pdf

Kenton, W. (2020, January 29). How Multiple Linear Regression Works. Retrieved May 17, 2020, from https://www.investopedia.com/terms/m/mlr.asp

Margalit, D., & Rabinoff, J. (n.d.). Interactive Linear Algebra. Retrieved from https://textbooks.math.gatech.edu/ila/least-squares.html

Spence, L. E., Insel, A. J., & Friedberg, S. H. (2019). *Elementary linear algebra: a matrix approach*. Noida, Uttar Pradesh, India: Pearson India Education Services.