Project Milestone 3 – Data Analysis and Regression

Instructions

1. Read this document carefully. You are responsible for following all instructions in this document.
2. Read the Learning Objectives at the end of the document to understand how your work will be graded.
3. Use professional language in all written responses and format all plots for technical presentation. See EPS01 and EPS02 for guidelines.
4. Good programming standards apply to all m-files.
5. Submit deliverables to Gradescope and to Blackboard. Name your files to match the format in the table below, where *SSS_TT* is your section and team ID (e.g., 001_03 is Section 001, Team 3)

Item	Deliverables
M3 Answer Sheet	Project_M3_AnswerSheet_*SSS_TT*.docx
M3 Algorithm	M3_Algorithm_*SSS_TT*.m
M3 Regression model	M3_Regression_*SSS_TT*.m
M3 Executive Function	M3_exec_*SSS_TT*.m
Gradescope Submission	M3_*SSS_TT*.pdf

See submission requirements on the last page of this answer sheet.

6. Complete the Assignment Header before starting the answer sheet.

Assignment Header

Section and Team ID (SSS_TT):	001-30

Team Member Name	Purdue Career Account Login
Alex Norkus	anorkus
Vincent Lin	lin971
Julius Mesa	jmesa
Surya Manikhandan	smanikha

Part 0: M2 Feedback Review

Reflect on your M2 feedback for the purpose of improvement. Your reflection should provide a clear, useful summary of your M2 feedback and provide a clear and practical plan to address the issues. Complete table 1 below.

Table 1. Feedback summary and plan

Part A: Summarize the feedback you received on M2 that could lead to improvements in your work.
The response from M2 states that we calculated the incorrect SSE values for the “clean” and “noisy” data. The correct data points were given with the feedback. Other feedback included incorrect naming standards for the user-defined functions and that the program does not run properly.

Part B: Explain how you will incorporate the M2 feedback to improve your parameter identification (do not just reword your response from Part A).

Based on the feedback received from M2, we can incorporate a better process for finding the SSE by backtracking where we went wrong in the M2 code and then comparing our new outputs with the answers given in the feedback. We will have a focus on the noisy data however, because the data we will be using from here on solely consists of noisy data.

After investigating our code, we were able to find why the SSE values were marked incorrect. Instead of calculating the SSE between the expected michaelis-menten curve and the points we plotted, the SSE provided on the M2 document were between the given (expected) values for v0 and the algorithm outputs, which is clearly incorrect. We will fix this issue in M3 by making sure the SSE value is calculated between the expected curve and the data points produced by the algorithm as was intended originally. Lastly, we will be more careful to follow programming standards by peer checking our work more carefully.

Part 1: Fully-Automated Parameter Identification

In Milestone 2, you developed two algorithms for identifying the enzyme parameters v_{0_i}, V_{max}, and K_m. You compared your algorithms' results to a known enzyme and identified potential improvements to your algorithms. **You must now select one algorithm to use for this milestone**. This user-defined function should include the better-performing elements of your two algorithms from M2 and incorporate the improvements you recommended in M2, Part 3. Name the updated algorithm **M3_Algorithm_SSS_tt.m**. The algorithm function must have appropriate inputs and outputs.

Next, develop an executive function to analyze the 100 test results of product concentration ([P]) provided by NovelEnzymes, Inc. in a **fully automated** way, and identify v_{0_i} for each enzyme test as well as the parameters V_{max} and K_m from the resulting Michaelis-Menten plot for each enzyme. Remember that the data provided to you represent five different enzymes: (i) 10 primary experiments per enzyme and (ii) 10 duplicates experiments per enzyme. Your executive function should call M3_algorithm_SSS_tt.m to process the data provided by NovelEnzymes.

You must develop a plan for your executive function before you start coding. Outline your plan for automating parameter identification using pseudocode (i.e., plain English text, not MATLAB code). Remember, it is valuable to develop and organize your programming ideas and solutions *before* you code. A well-developed plan reduces coding frustrations. Complete your plan in Table 2 below.

Table 2. Executive function plans

[redacted]
First off, our executive function needs to cut off the dataset to an appropriate length and trim the time series accordingly. This is necessary because the columns have wildly varying sizes and this leaves in a lot of NaN values which need to be eliminated from the dataset before any calculations can be made. This can be done by using the "isnan" and redirection operators to appropriately trim the dataset without the need for any looping through the array. The executive function will then call the Algorithm (M2_Algorithm_001_30). First, the Algorithm calls the smoothing function (M3_Smooth_001_30) in order to eliminate as much noise, variability, and bias in the dataset as possible. For any given enzyme, this process is repeated 10 times at each of the

substrate concentrations. Then, the algorithm uses the slope equation rise/run to find V_0 for each of those 10 experiments and returns these values to the executive function.

Using the Hannes-Wolf method we discussed in detail in the previous milestone document, we can linearize our reaction velocity plot, which allows us to calculate the V_{max} value of the given enzyme and the K_m value through the use of the linear slope and intercept model. The executive function will automatically repeat this process 5 times for each enzyme. After the parameters are identified, the exec function will calculate the expected reaction velocity curve using the michales-menten equation which allows us to calculate SSE. Finally, the exec function will plot the raw and linearized reaction velocity plots and their corresponding best fit lines for user reference.

The final output to the command window is a vmax array containing all the vmax values for the 5 enzymes, the Km array containing the km values for all 5 of the enzymes, and the SSE array which was calculated by the exec function at the very end.

We believe our exec function is a great solution as it is completely automated. It only takes 1 function call for all the values to be produced and no other user intervention is needed. Example execution and outputs can be found in Part 2 in figures 3.1-3.6

After you complete your plan, translate it into an executive function. Name the executive function **M3_exec_SSS_tt.m**.

Part 2: Goodness of Fit

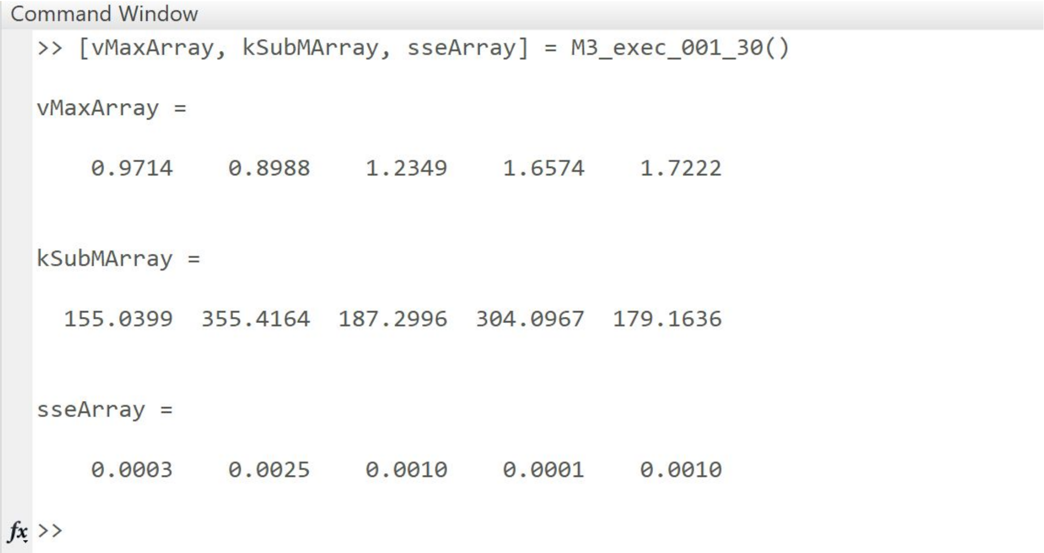
The next step is to examine how well your parameters fit the Michaelis-Menten model. In your executive function, add the calculations for SSE for the Michaelis-Menten model for each of the 5 enzymes. Complete Table 3; be careful to use appropriate decimal places.

Table 3. Enzyme parameters and model goodness of fit

Numbers are kept at 4 decimal places for purposes of making accurate comparisons to benchmark performance of our algorithm with others.

Enzyme	Enzyme Parameters		SSE
	V_{max} (µM/s)	K_m (µM)	$(\mu M/s)^2$
NextGen-A	0.9174	155.0399	.0003
NextGen-B	0.8988	355.4164	.0025
NextGen-C	1.2349	187.2996	.0010
NextGen-D	1.6574	304.0967	.0001
NextGen-E	1.7222	179.1636	.0010

Figure 3.1 Executive Function Command Window Output For Milestone 3

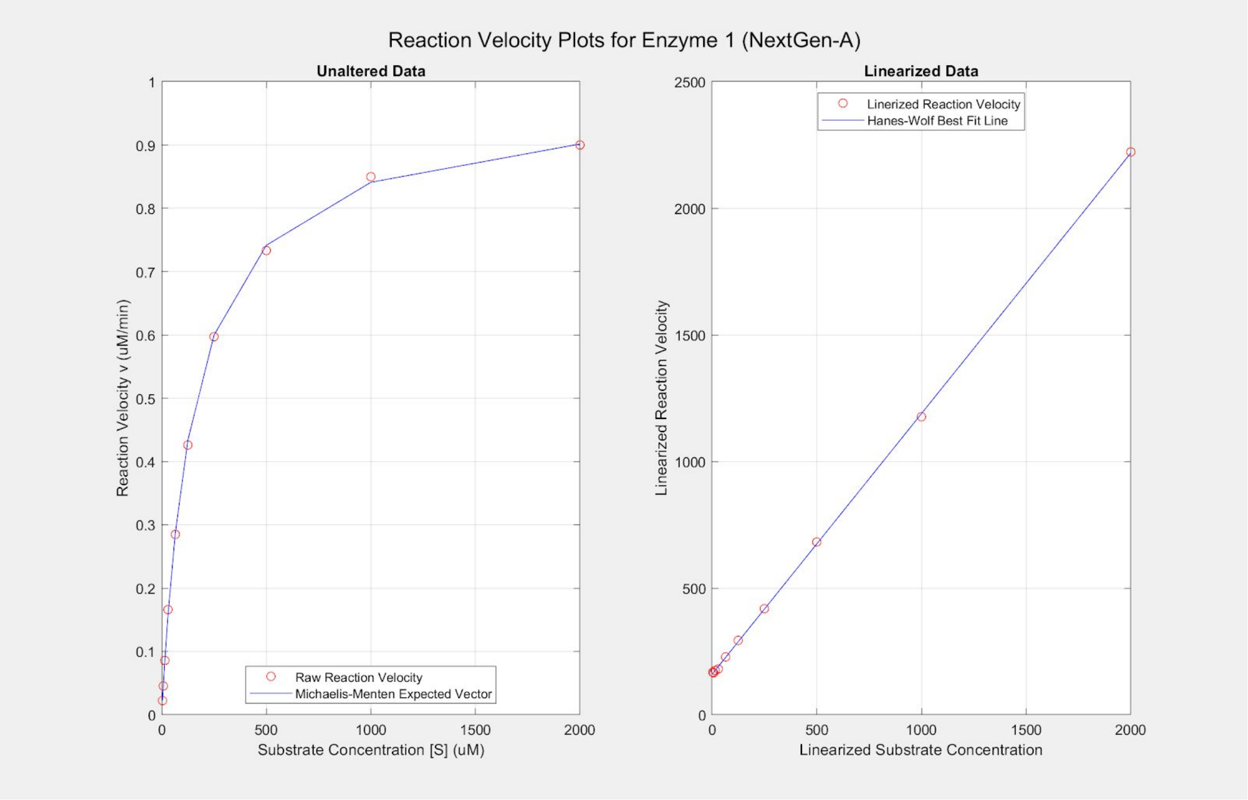


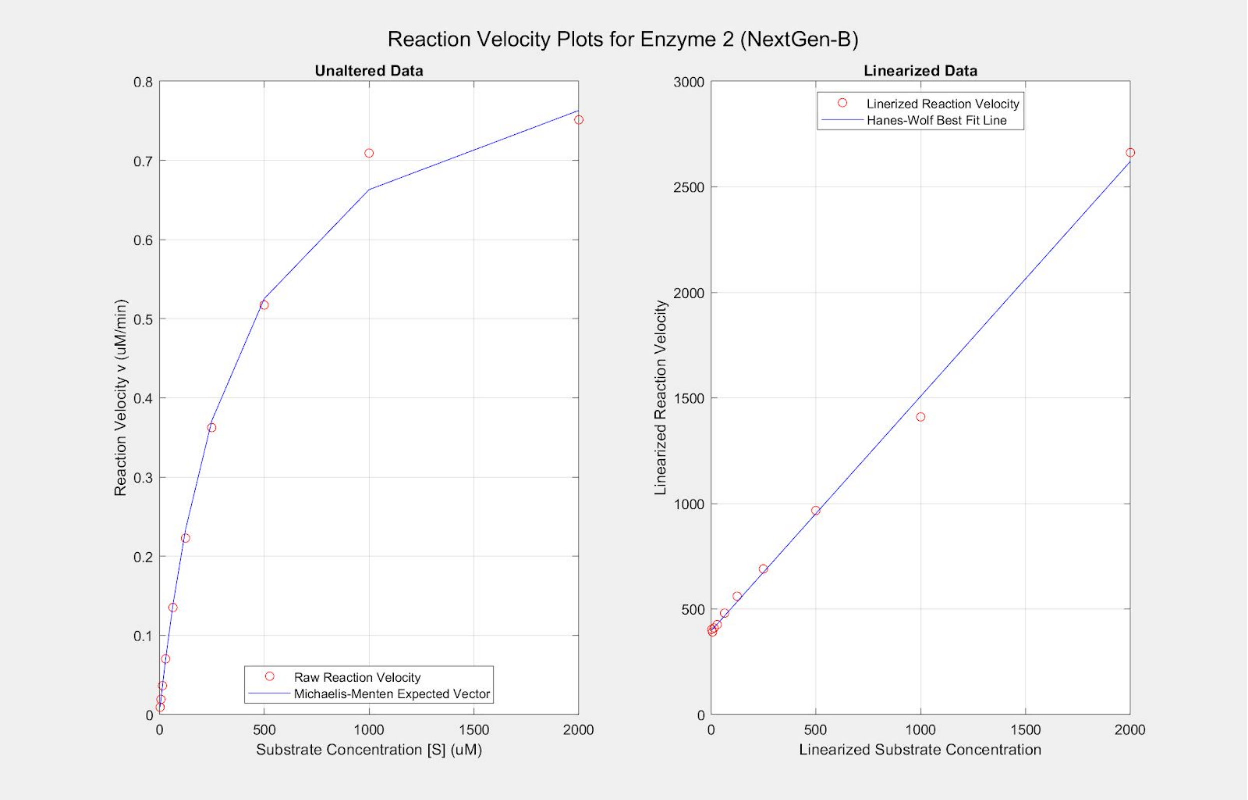
```
>> [vMaxArray, kSubMArray, sseArray] = M3_exec_001_30()

vMaxArray =

    0.9714    0.8988    1.2349    1.6574    1.7222


kSubMArray =

  155.0399  355.4164  187.2996  304.0967  179.1636


sseArray =

    0.0003    0.0025    0.0010    0.0001    0.0010

fx >>
```

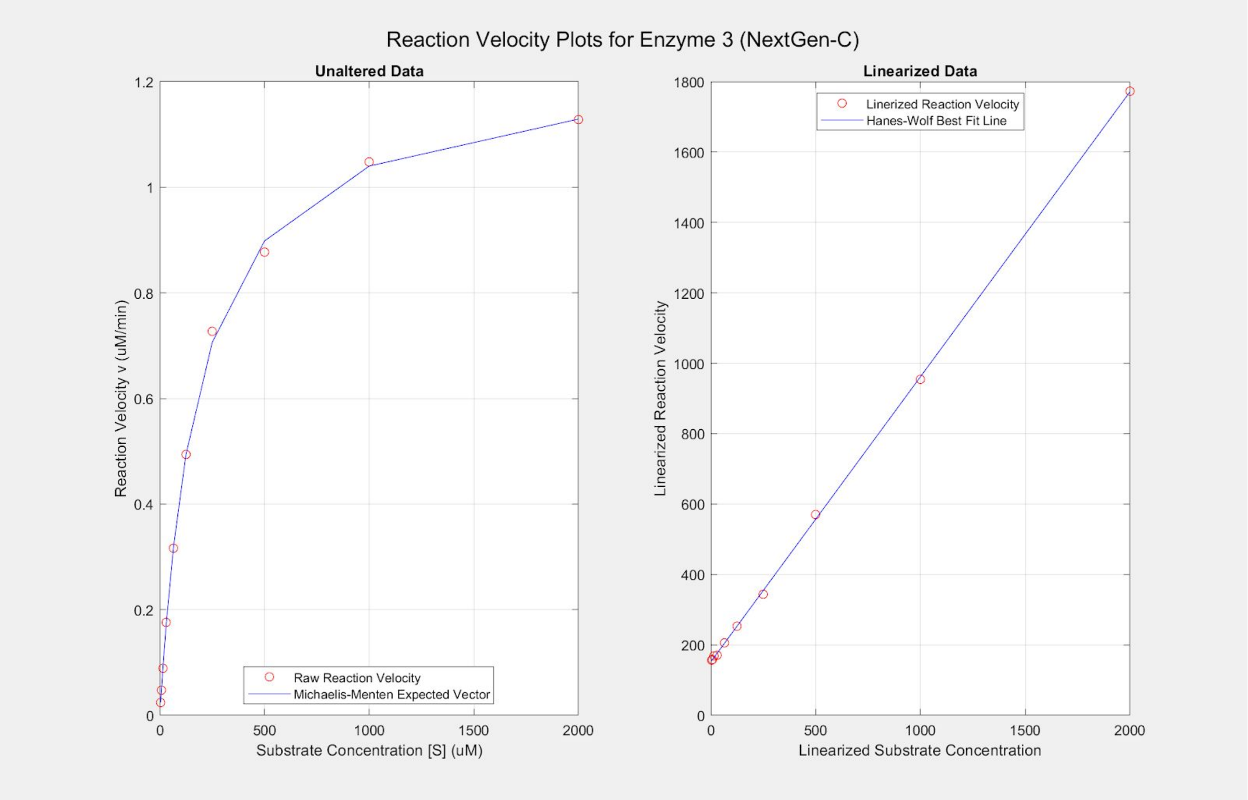
Examine the SSE results for each enzyme. Evaluate whether or not the parameters identified by your algorithm used in the Michaelis-Menten model fit your data. Reflect on whether or not you need to make changes to your M3 algorithm before you submit this milestone.

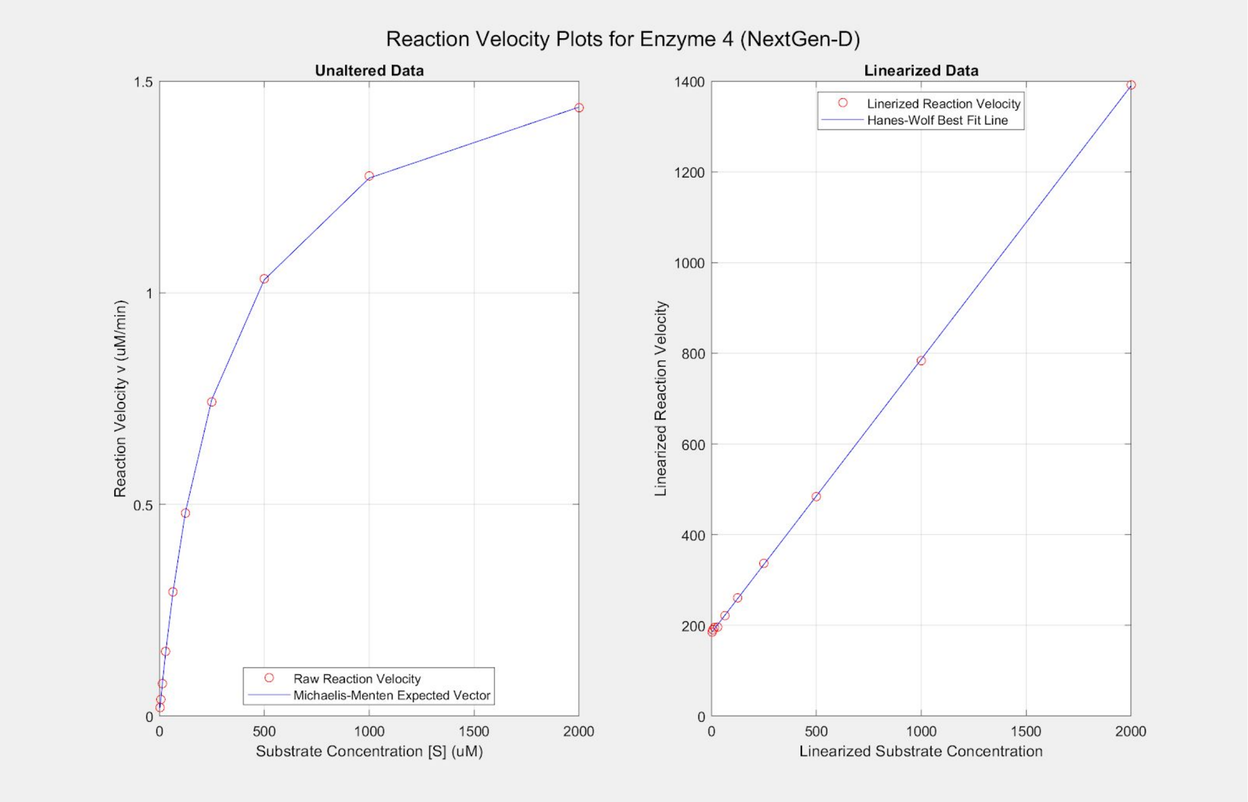
Looking at the SSE values, they are very small meaning our expected michaelis-menten curve and our algorithm's data points are very close together and therefore have very low error. This indicates a near-perfect fit, a finding which is further supported by a visual inspection of out plot outputs (Figures 3.2-3.6). Therefore, we are confident no changes need to be made at this point in time.

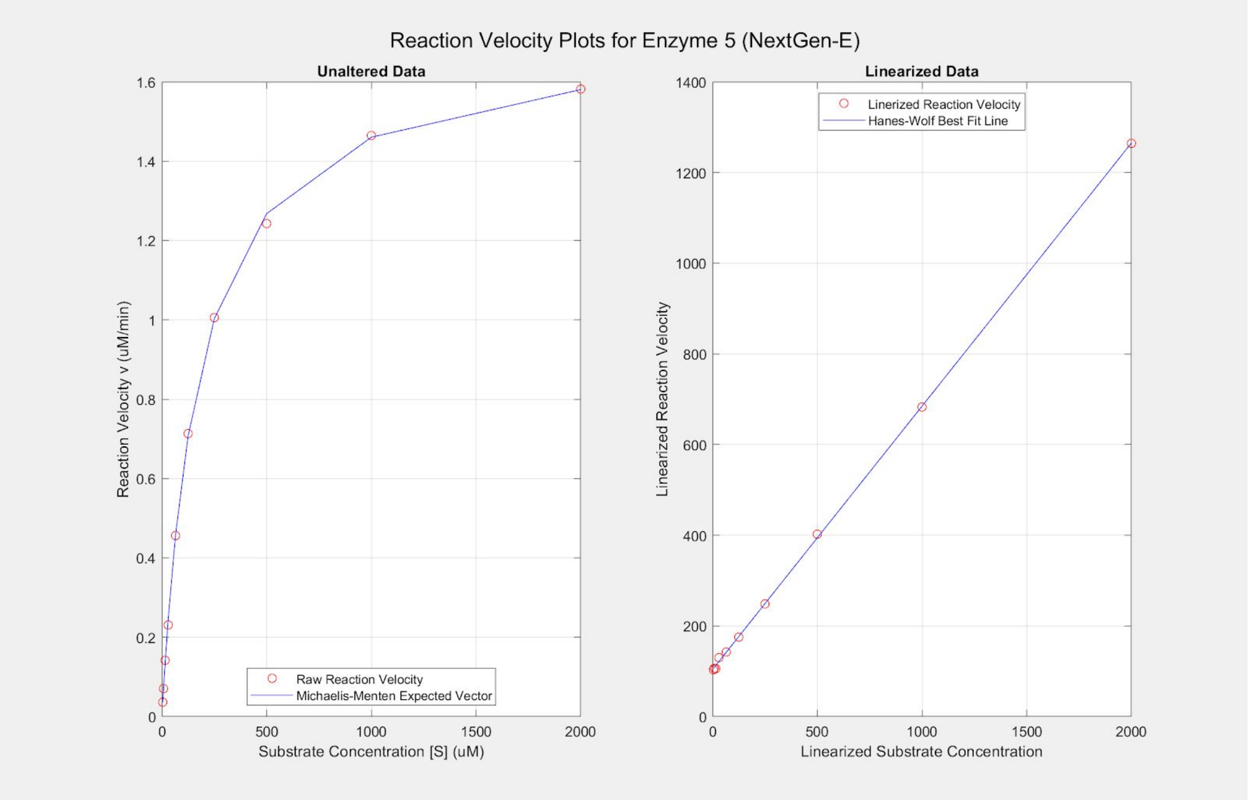
Figures 3.2 - 3.6 Executive Function Reaction Velocity Plot Displays






Reaction Velocity Plots for Enzyme 3 (NextGen-C)
Unaltered Data
Reaction Velocity v (uM/min)
1.2
1
0.8
0.6
0.4
0.2
0
0
500
1000
1500
2000
Substrate Concentration [S] (uM)
Raw Reaction Velocity
Michaelis-Menten Expected Vector
Linearized Data
Linearized Reaction Velocity
1800
1600
1400
1200
1000
800
600
400
200
0
Linearized Substrate Concentration
Linerized Reaction Velocity
Hanes-Wolf Best Fit Line


Reaction Velocity Plots for Enzyme 4 (NextGen-D)
Unaltered Data
Reaction Velocity v (uM/min)
1.5
1
0.5
0
0
500
1000
1500
2000
Substrate Concentration [S] (uM)
Raw Reaction Velocity
Michaelis-Menten Expected Vector
Linearized Data
Linearized Reaction Velocity
1400
1200
1000
800
600
400
200
0
Linearized Substrate Concentration
Linerized Reaction Velocity
Hanes-Wolf Best Fit Line



Part 3: Price Regression

Write a user-defined function called M3_Regression_SSS_tt.m that determines a regression model for NovelEnzymes' price data as a function of enzyme performance; consult the M3 memo for more information. The regression user-defined function must be called by the executive function, with appropriate inputs and outputs being passed between the functions. You will need to carefully consider the shape of the relationship between price and performance. What type of function best models this relationship?

The regression user-defined function must generate a plot showing the pricing regression results: plot price versus Michaelis constant for all given pricing data, and overlay your regression model (in the generalized equation form) on the same figure. The regression equation must be displayed on the plot in a suitable location.

Report the metrics for your linearized regression model in Table 4 below.

Type of function	**The power function represents the best for the relationship between price and enzyme performance.**
Concern:	**However, when the Michaelis Constant (uM) reaches 0, meaning the enzyme performs perfect, the price is close to infinity. We will have to deliberate more as a team when it comes to extrapolation of the data**

Table 4. Regression metrics for linearized price versus performance data.

Parameter	Value
SSE ($\$^2$)	3.549e+04
SST ($\$^2$)	8.2534e+05
r^2	0.9570

Part 3: Observations and Improvements

In M2, you were asked to generate ideas about how to improve your algorithm for parameter identification. In M3, you have improved your algorithm and applied your algorithm to 100 different test results, and seen its performance across all that data. Based upon your M3 work, again consider potentially useful improvements and provide two specific ways you can improve your parameter identification model. Briefly explain each suggestion using evidence-based rationales. You do not need to code these changes; at this point, simply describe changes you think might be useful.

Be sure to:

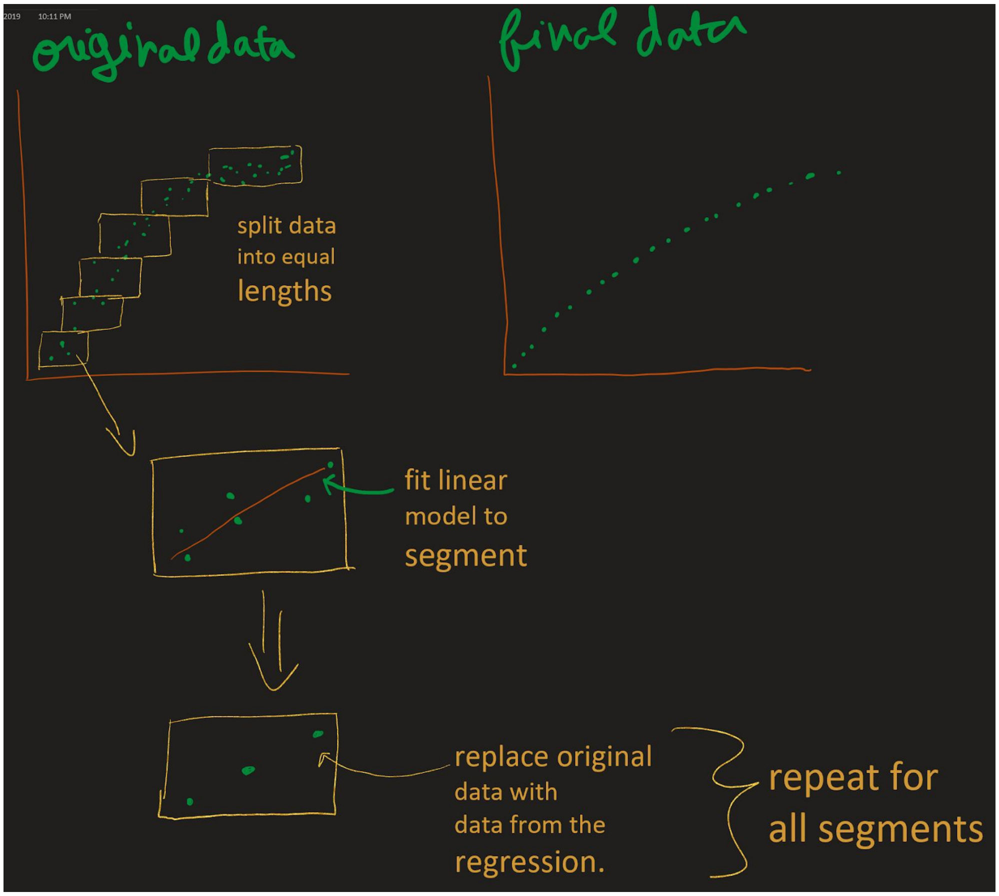
- explain which parameter(s) your improvement will target,
- explain the improvement with a level of detail that can be understood by others (provide sketches or flowcharts as necessary to clarify your improvement),
- describe the performance metrics you will use to determine whether your proposed improvement really does improve your solution, and
- provide evidence-based rationales for each proposed improvement and the metrics selected. Your rationales should answer the questions:
    - What is your evidence that this improvement is necessary?
    - Why is this method for making the improvement a good idea - what is your evidence?
    - Why is this metric a good idea - what is your evidence?

Discuss improvements that resolve issues that you see in your current algorithm in Table 5 below. Cite any external references in Table 6.

Table 5. Algorithm improvements (add improvement blocks as needed)

Improvement 1. Parameter(s) Targeted: Smoothing Function
Description *We believe overhauling the smoothing function would significantly improve final results for v0, Vmax, and Km, especially given that with noisy data which sometimes possess significant bias. Currently, we are using the moving-average method, which uses a shifting interval and replaces the data points in that interval with the average. While this works well, a significant disadvantage is the loss of data points. Instead, we would like to use a method which results in no data loss.Our new function will use the LOESS method of smoothing. Essentially, it stands for Local Regression Smoothing, and works similarly to the moving average (Jacoby, 2000). Specifically, the LOESS algorithm takes a segment of data of a given width and finds the regression of those data points with any power function (we will*

likely use linear). then, the algorithm will replace the points in that segment with an equal number of data points from the regression model. This is repeated for all segments. A simple explanation diagram from one of our brainstorming sessions can be found below:


2019 10:11 PM
original data
final data
split data into equal lengths
fit linear model to segment
replace original data with data from the regression.
repeat for all segments

Metrics to Determine Improvement

We will be comparing the SSE of the reaction velocity data produced through the use of the current algorithm and once again through the use of the proposed LOESS function. We will deem the function with the lowest final SSE values the winner.

Rationale for Improvement and Metrics

The way our current smoothing function works allows it to be significantly affected by the large data bias as the moving average method works best on long-term trends as opposed to shorter timescales. Since we are only going to be dealing with the first few points for out v0 calculations, a better algorithm is likely necessary. The new approach to the smoothing function would allow us to see the bigger picture since there is minimal loss of data, bringing the points closer to one smooth curve rather than just averaging graph points.

Our metrics to determine improvements are valid as smaller SSE values indicate that the smoothing function has done a better job of conditioning the data for calculation, and therefore was able to being the calculated values closer to the expected vector form the michaelis-menten equation. Plus, it is an easy metric to determine so it requires little modification.

Improvement 2. Parameter(s) Targeted: Handling of original and duplicate test data

Description

Initially, we planned to average the values for the original and duplicate test before smoothing and calculating the v0 values in hopes of making a significantly more accurate calculation. We thought this was a good idea because it was yet another way to get rid of noise in the datasets. However, we ran into many technical issues implementing this idea since the vectors for the original and duplicate test were not always the same size. This caused multiple array dimension and out of bounds errors which, due to heavy time constraints, were unable to fix. Therefore, our M3 algorithm finds the v0 values for both the original and duplicate data and then averages those arrays (v0 arrays are always the same dimension of 10 elements). However, it is likely that this method compounds error in our algorithm, as indicted by some members of the ENGR 132 teaching team.

For the next milestone, we will strive to smooth our original test data and our duplicate data prior to smoothing them and performing parameter calculations.

Metrics to Determine Improvement

We will use the same SSE method described above in improvement 1 in order to see if the new smoothing function is able to smooth the data better and therefore pull our experimental points closest to the theoretical.

Rationale for Improvement and Metrics

averaging the v0 values for the original and duplicate at the very end compounds the error of our algorithm, which is not an ideal situation. As we described in detail at the “description” section, it is likely that averaging the original and duplicate test data first will significantly improve our curve’s smoothness as it will help eliminate a lot of the spikes present within the dataset which skew our average-based calculation.

Table 6. References used in evidence-based rationales

Jacoby, William G. “Loess: a Nonparametric, Graphical Tool for Depicting Relationships between Variables.” *Electoral Studies*, vol. 19, no. 4, 2000, pp. 577–613., doi:10.1016/s0261-3794(99)00028-1.

[redacted]

1. Rename this answer sheet to be **Project_M3_AnswerSheet_*SSS_TT*.docx** where ***SSS*** is your section number (e.g., 001 for section 001) and ***TT*** is your team number (e.g., 07 for team 7). Save the answer sheet as a PDF.
2. Publish each function you created in this milestone. Note: you may publish sub-functions (such as M3_Algorithm_SSS_tt.m) without filling in appropriate inputs or outputs – this will generate an error in the published code for that function that will not affect your grade. Note: these functions must run properly when called by your executive function
3. Merge the answer sheet PDF and the published code PDFs into one PDF file named **M3_*SSS_TT*.pdf**.
4. Select one person to submit the PDF for the team. That person should
    a. Log into Gradescope and submit **M3_*SSS_TT*.pdf** to the **M3** assignment.
    b. Indicate which pages correspond to each part of the milestone.

       Failure to tag the items appropriately will result in your work receiving losing credit.
    c. Select all team members for the group assignment.
    d. Double-check that all team members are assigned to the submission.
5. Select one person to submit all m-files, data files, and answer sheet to the **M3 dropbox** on Blackboard.

   Failure to submit your files to Blackboard may result in every team member receiving a zero on this assignment.
6. Each team member should confirm that they are part of the submission and that all parts of the answer sheet were properly tagged.
7. After submission, distribute the submitted files to all team members. *Ensure all members of the team have copies of the submitted files.*

[redacted]

Process Awareness (PA)
Reflect on both personal and team's problem solving/design approach and process for the purpose of continuous improvement.
PA02. Identify limitations in the approach used.
PA03. Identify potential behaviors to improve approach in future problem solving/design projects.

Idea Fluency (IF)
Generate ideas fluently. Take risks when necessary.
IF03. Generate testable prototypes (including process steps) for a set of potential solutions.

Evidence-Based Decision Making (EB)

Use evidence to develop and optimize solution. Evaluate solutions, test and optimize chosen solution based on evidence.

EB01. Test prototypes and analyze results to inform comparison of alternative solutions.

EB03. Clearly articulate reasons for answers with explicit reference to data to justify decisions or to evaluate alternative solutions.

EB05. Present findings from iterative testing or optimization efforts used to further improve aspect or performance of a solution.

EB06. Clearly articulate reasons for answers when making decisions or evaluating alternative solutions.

Solution Quality (SQ)

Design final solution to be of high technical quality. Design final solution to meet client and user needs.

SQ01. Use accurate, scientific, mathematical, and/or technical concepts, units, and/or data in solutions.

Information Literacy (IL)

Seek, find, use and document appropriate and trustworthy information sources.

IL04. Include citations within the text (in-text citations) that show how the references at the end of the text are used as evidence to support decisions.

IL05. Format reference list of used sources that is traceable to original sources (APA or MLA are recommended)

Engineering Professional Skills

PC05. Fully address all parts of assignment by following instructions and completing all work.

EPS01. Use professional written and oral communication.

EPS02. Format plots for technical presentation.

Programming

MAT01. Develop code that follows good programming standards.

MAT03. Perform mathematical operations and calculations within MATLAB.

MAT05. Create and use MATLAB scripts and functions.

MAT08. Debug scripts and functions to ensure programs execute properly, perform all required tasks, and produce expected results.

MOD01. Create mathematical models to describe relationships between data.