



Tools for Artificial Intelligence with MATLAB, initiation (TAIM)

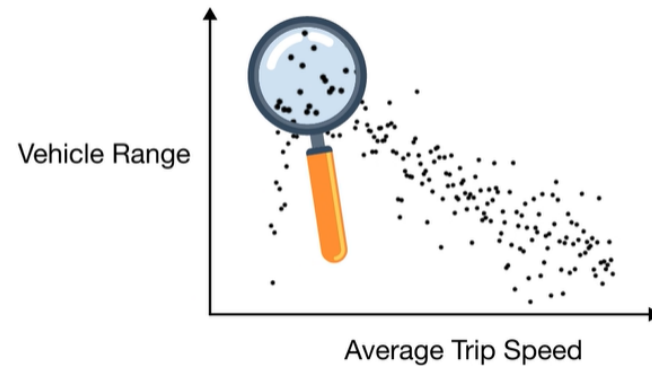
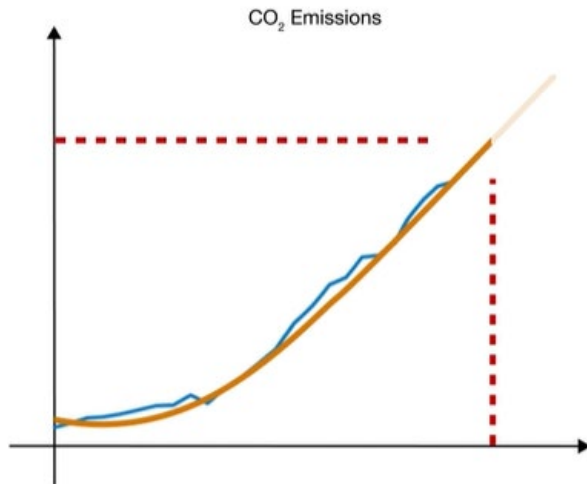
José Antonio Lázaro

Data analysis: 2 dimensions (Curve fitting)

Barcelona, 3, February 2025

What for?

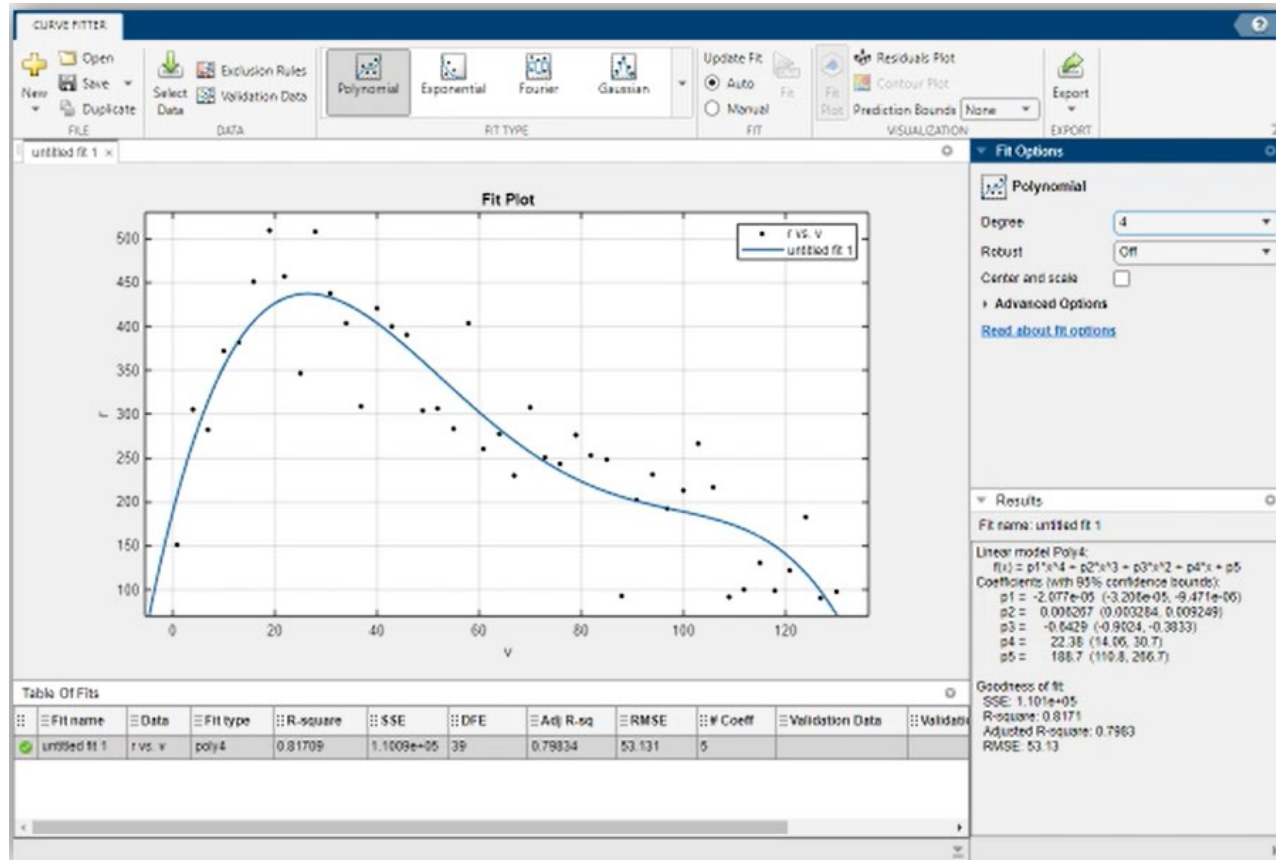
- E.g. Future predictions and estimations



- Optimization: e.g. the best speed of a vehicle to reach a maximum distance?

What for?

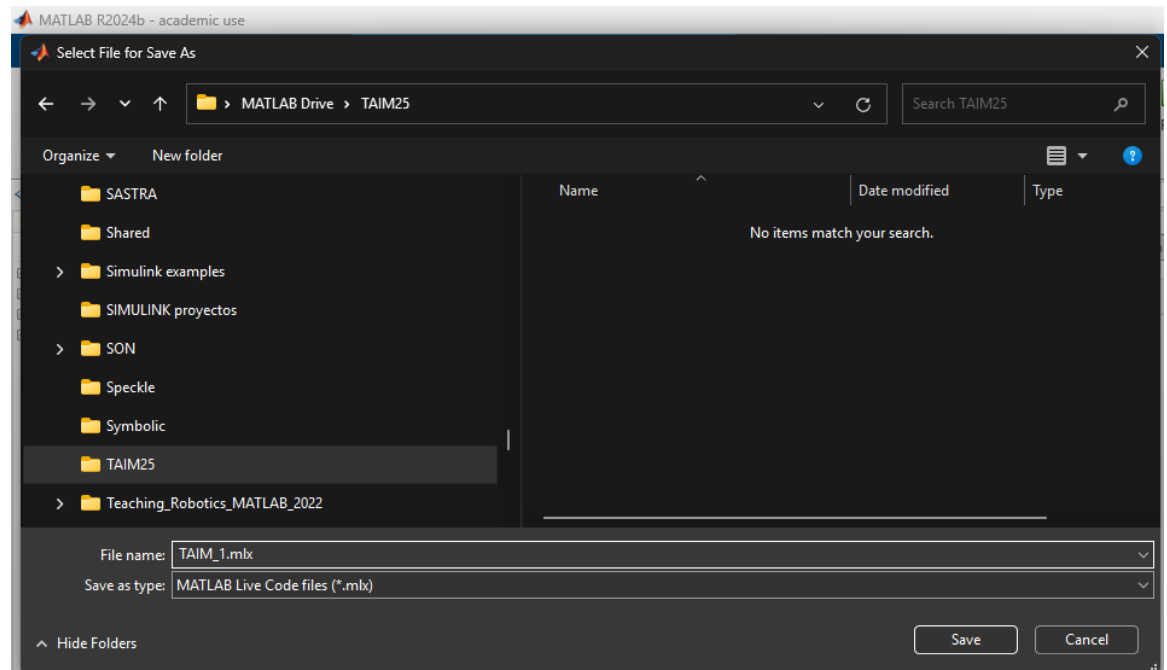
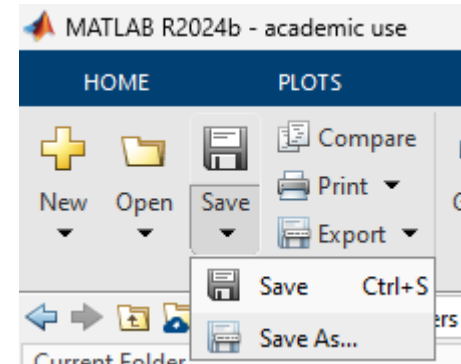
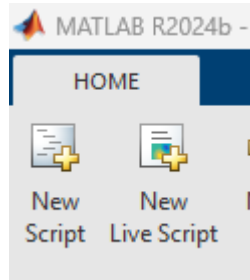
- Optimization: e.g. the best speed of a vehicle to reach a maximum distance for an electric vehicle



Let's go

Open Matlab

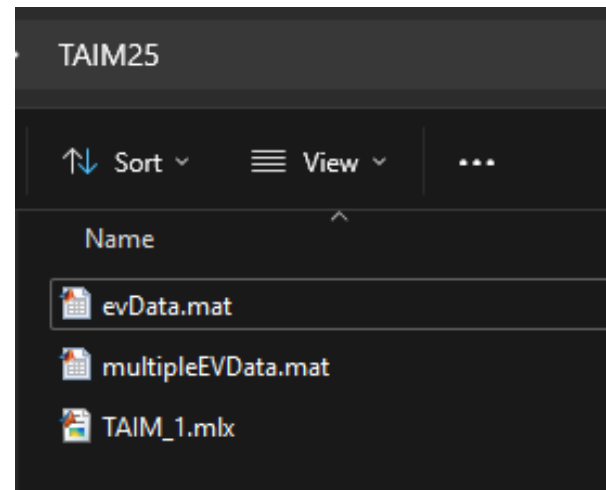
- Do “New Live Script”
- Go “Live Editor”
- And “Save as”
- Create a Folder for your course “TAIM25”
- Select a Name and save it. (E.g. “TAIM_1”)
- NO SPACES at the NAME



Let's go

Go to “My_TECH_SPACE”

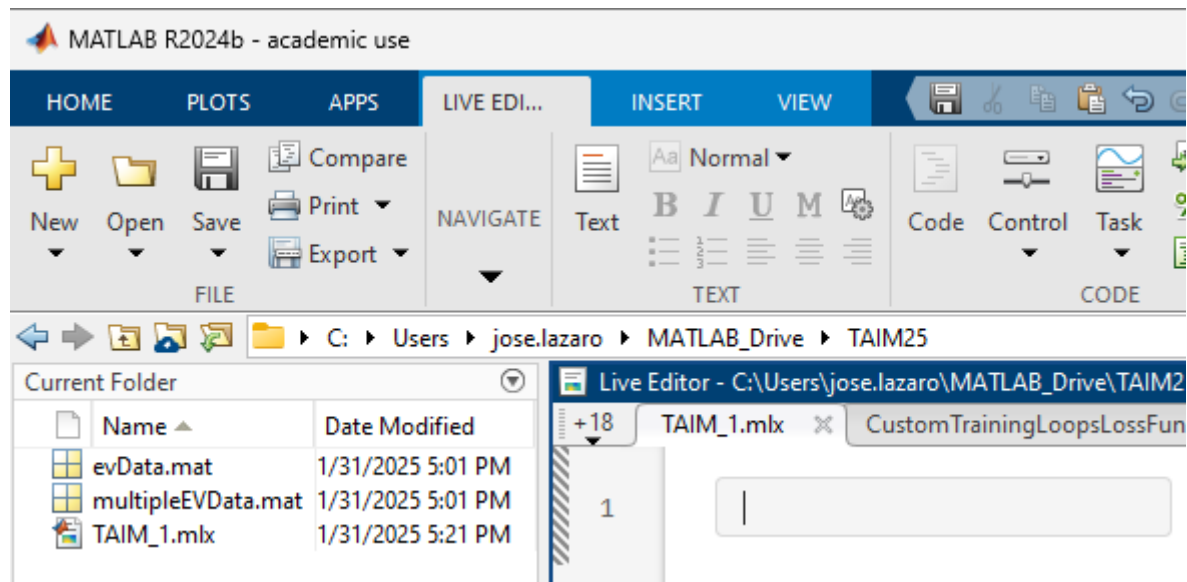
- Download the files: “evData.mat” and “multipleEVData.mat”
- Copy them at your “TAIM_1” folder.
- It should look like this



Let's go

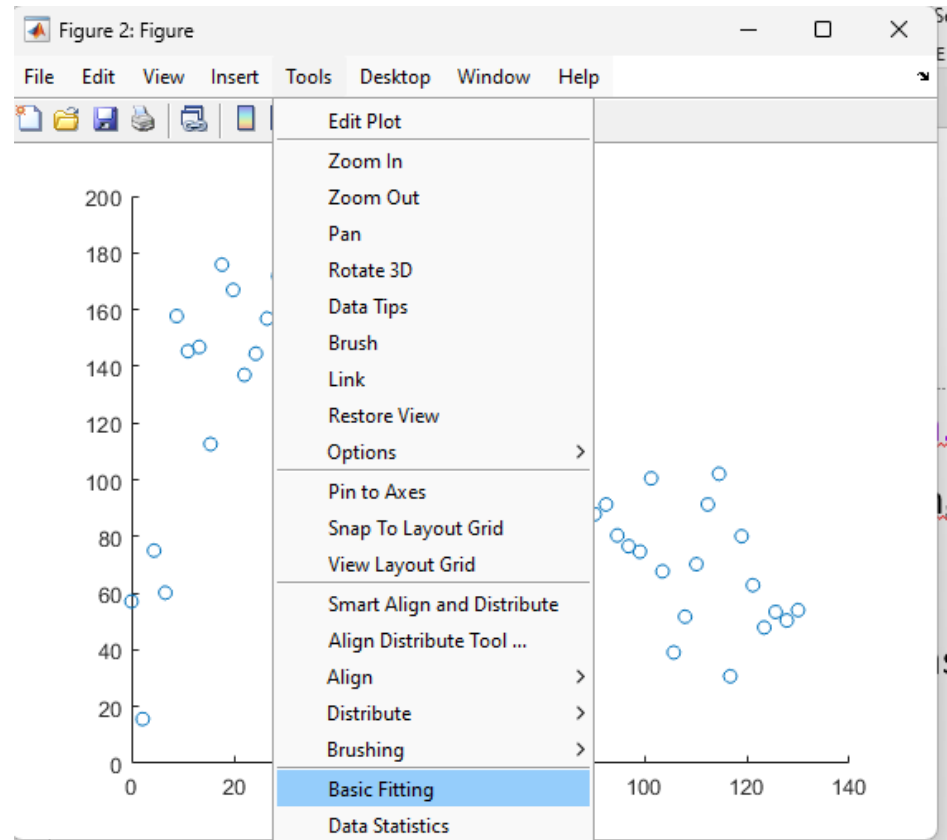
Go back to Matlab

Your file should look like this:



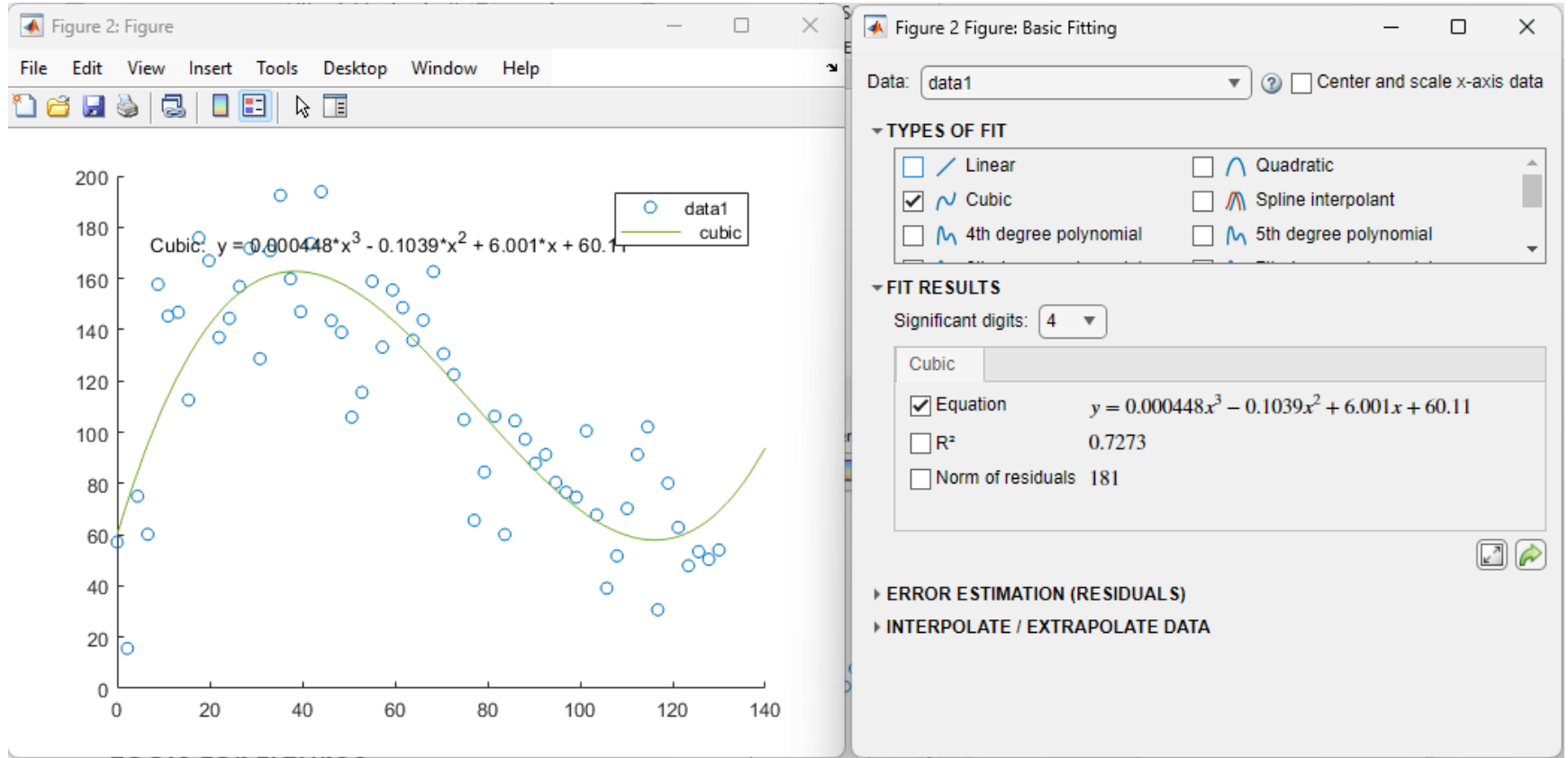
Let's go

- Start loading the working data: “load **evData.mat**”
- Now let's draw the data:
“scatter(speed,range)”
- Now we can do a 1st basic fitting suing the basic fitting tools for figures



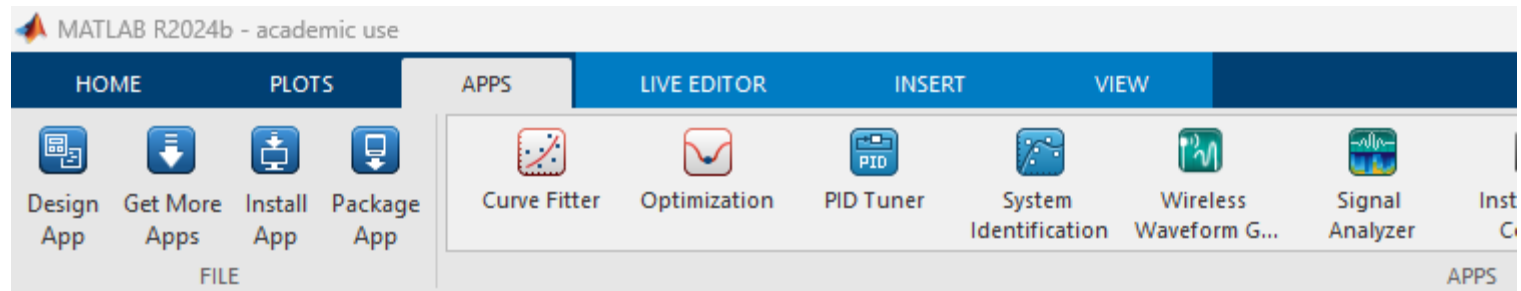
Let's go

- You can do a 1st fitting, though maybe too basic



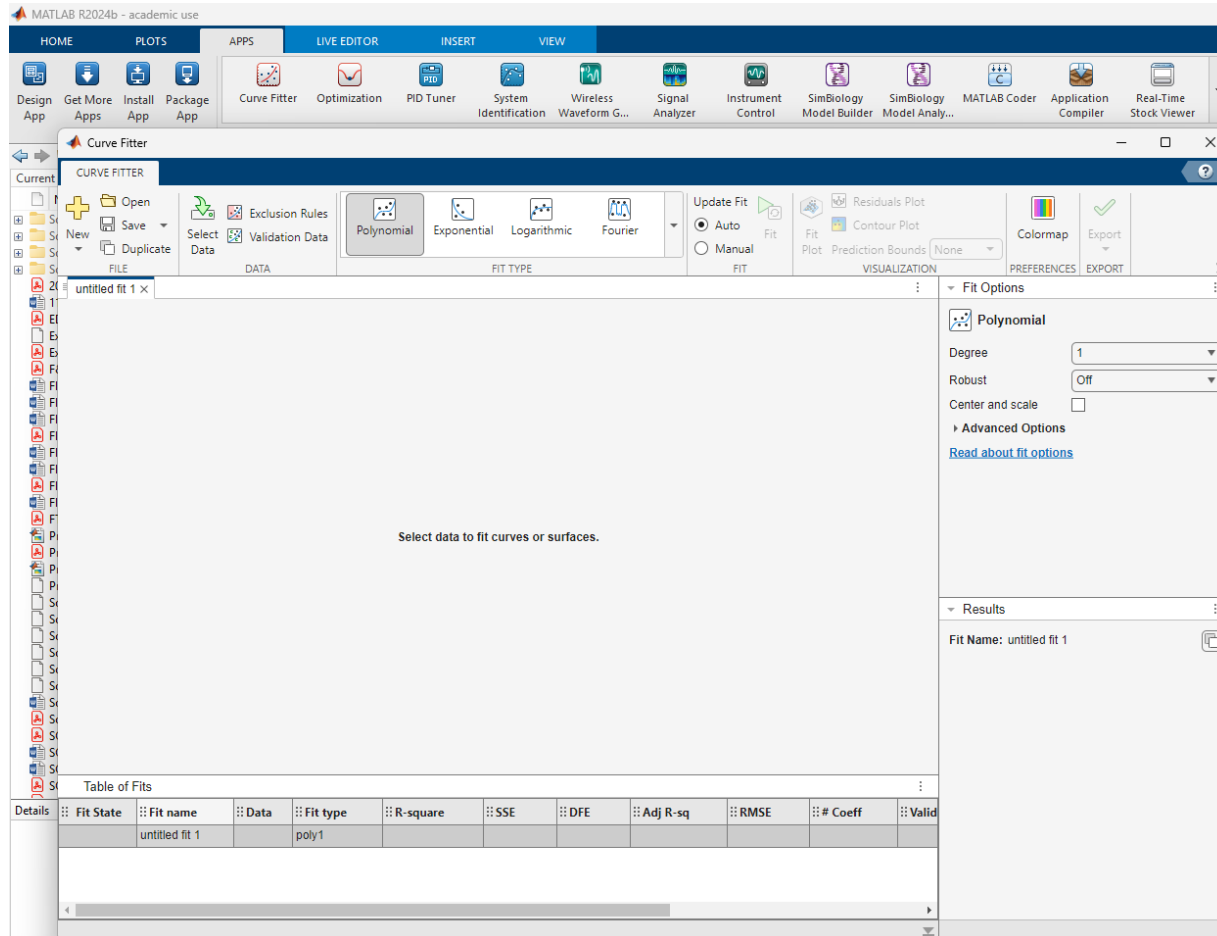
Let's do it better

- Go to “APPS”
- Open “Curve Fitter” Application



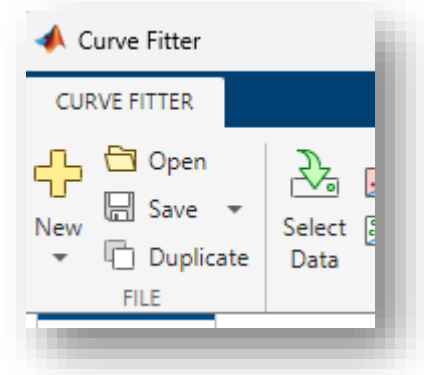
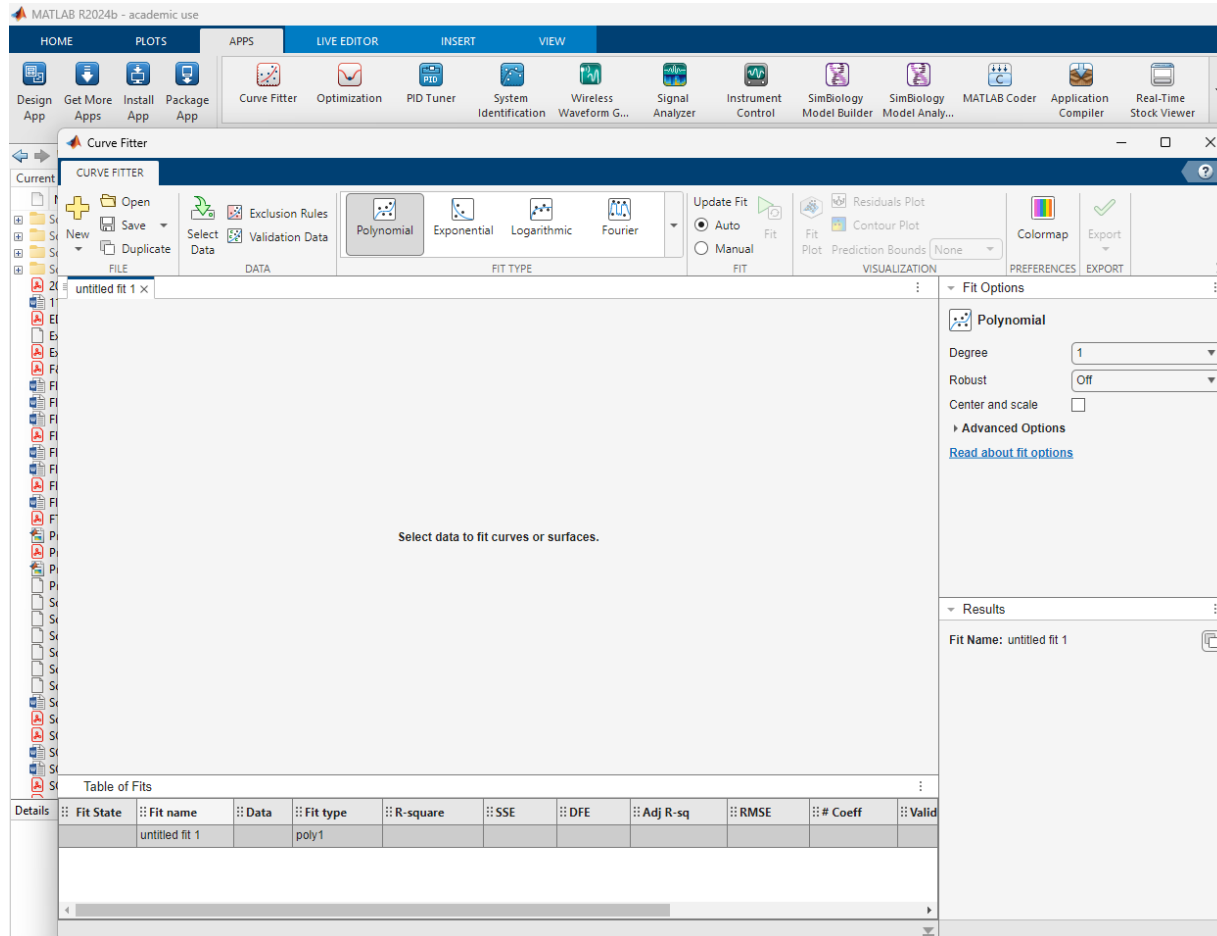
Curve Fitter App

- You get this



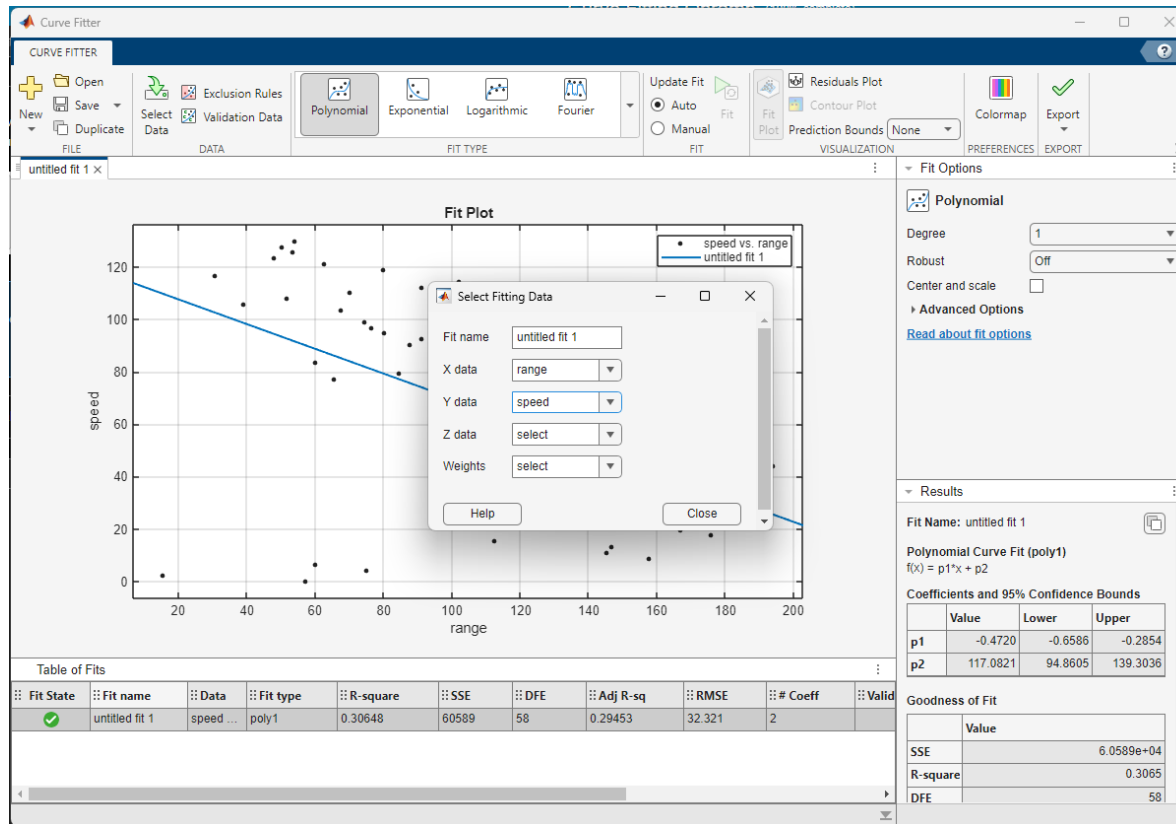
Curve Fitter App

- Import the Data at the new tool



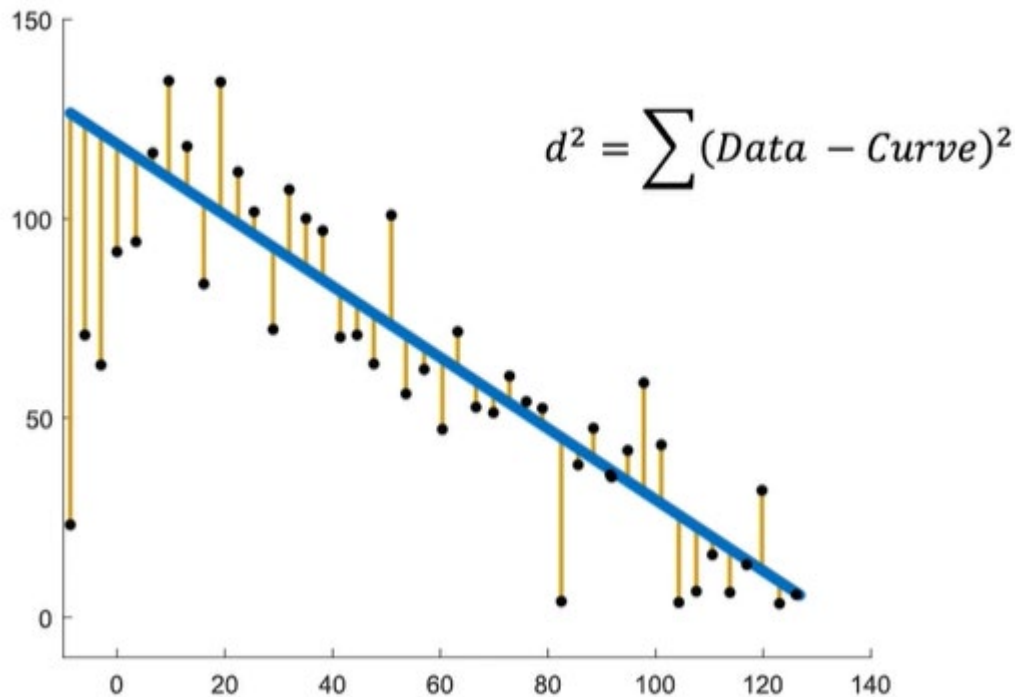
Curve Fitter App

- Import the Data at the new tool, selecting X = range & Y = speed



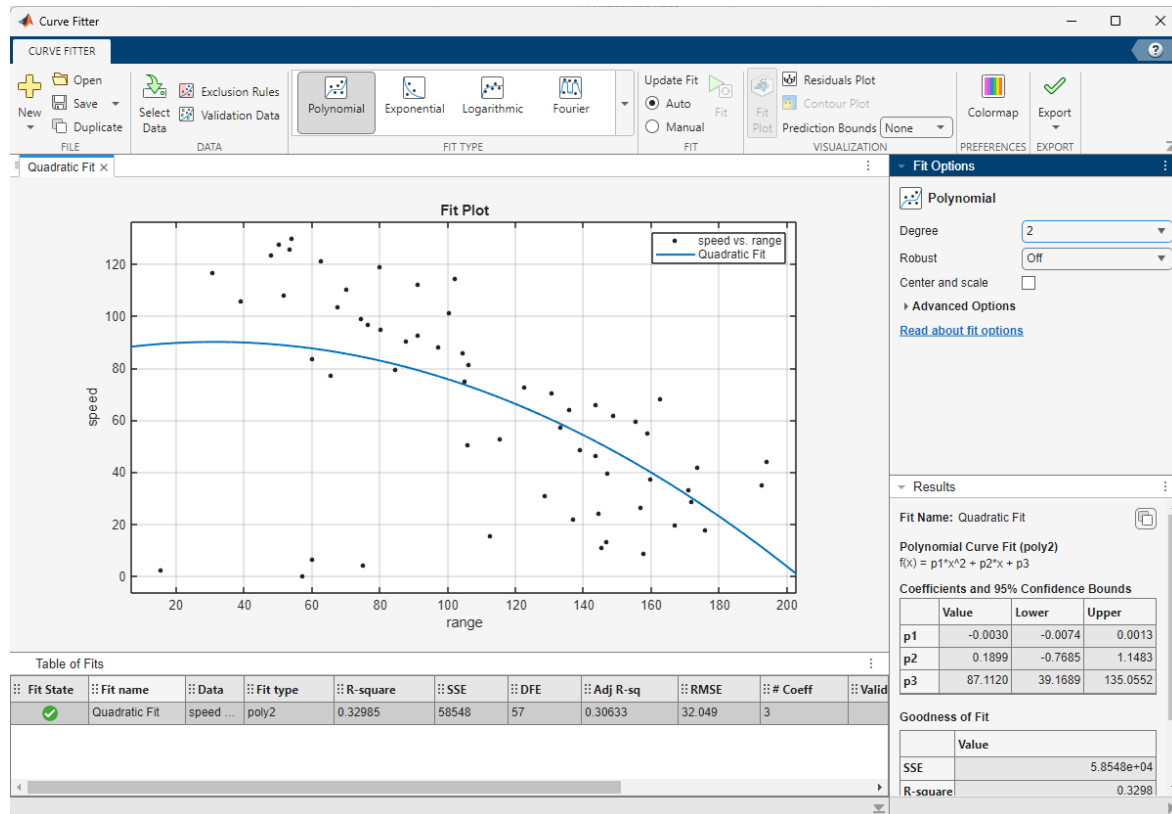
Curve Fitter App

- A 1st curve is done, but: Is this the best mathematical model of the data?
- The linear curve has been calculated, automatically to minimize the “d²”



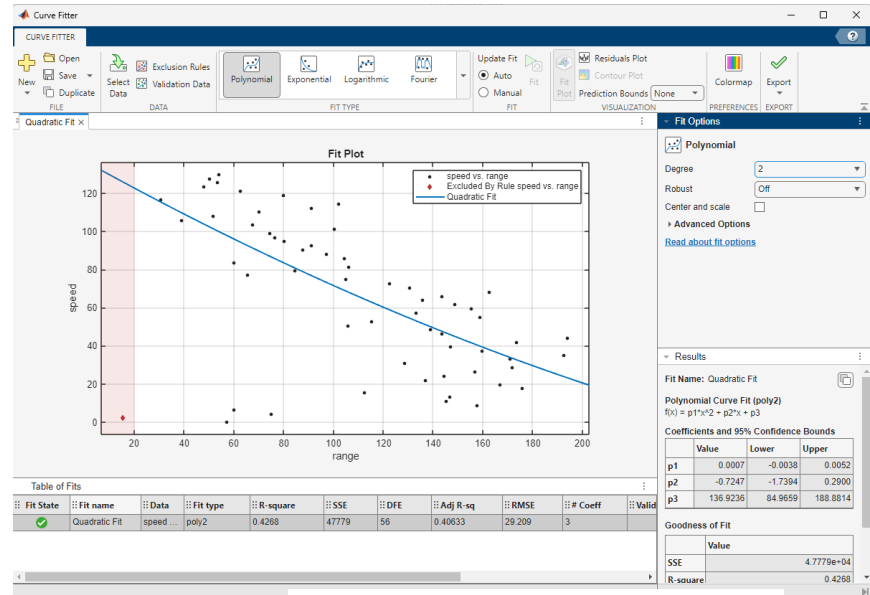
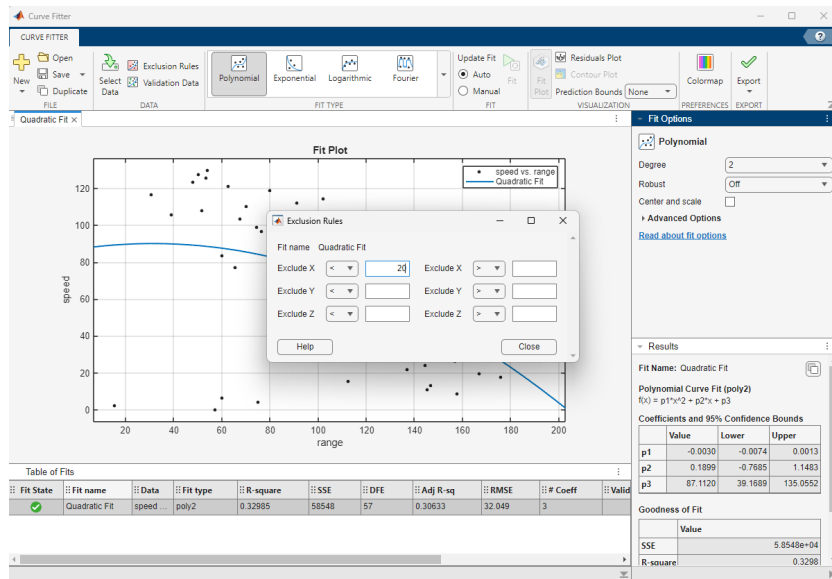
Curve Fitter App

- You can also try a quadratic fit

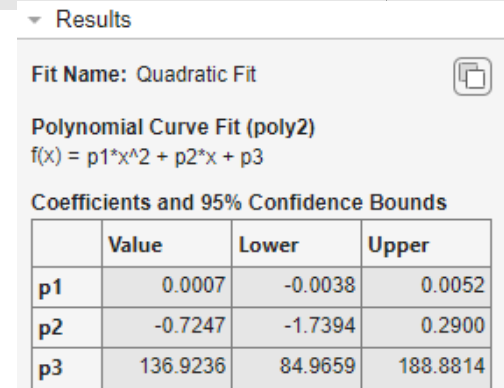


Curve Fitter App

- You can also exclude some data, if it is considered not relevant

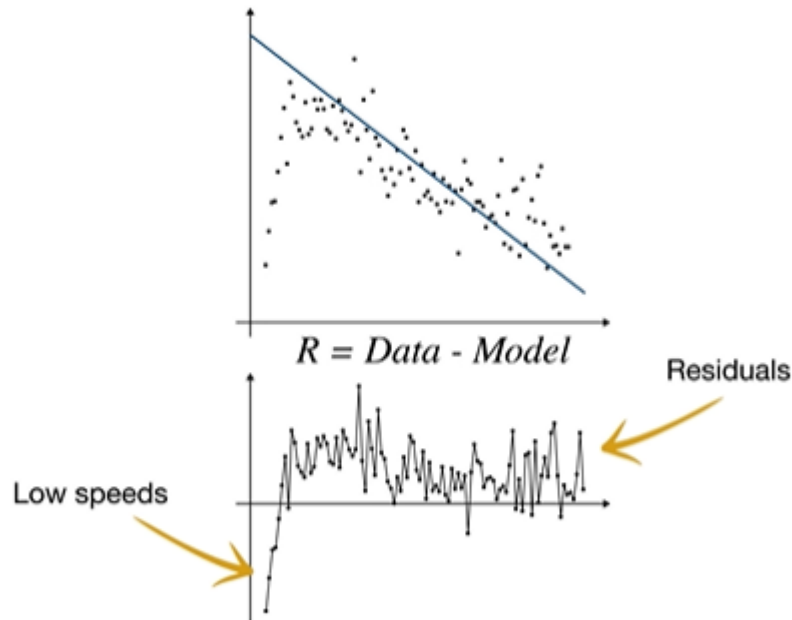


- But, then, the coefficient for X^2 is nearly zero -> This means that without this data, the rest of the data fits better to a linear curve.



Curve Fitter App

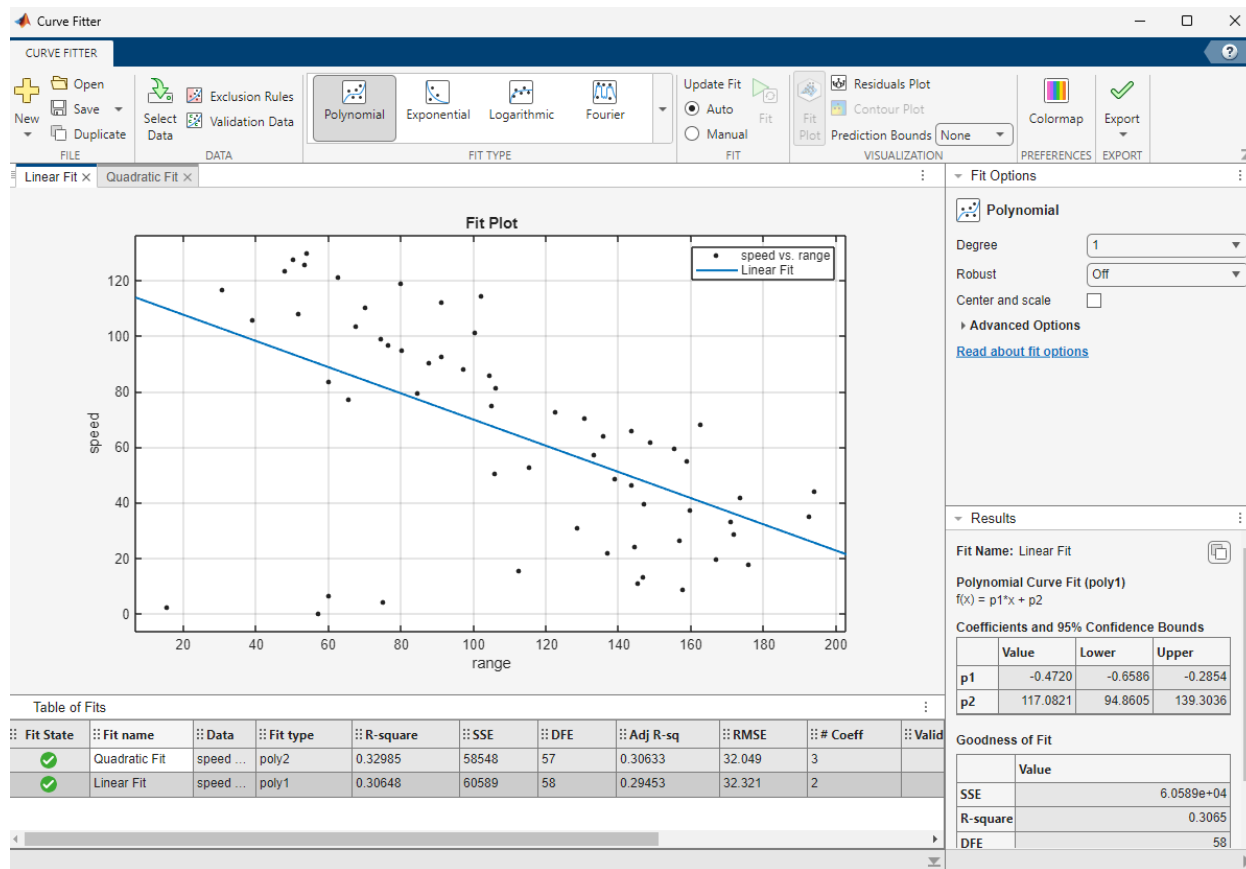
- A more detailed analysis can be done looking to the “residuals”



- For a linear curve, it says that the low speeds are “badly” represented.

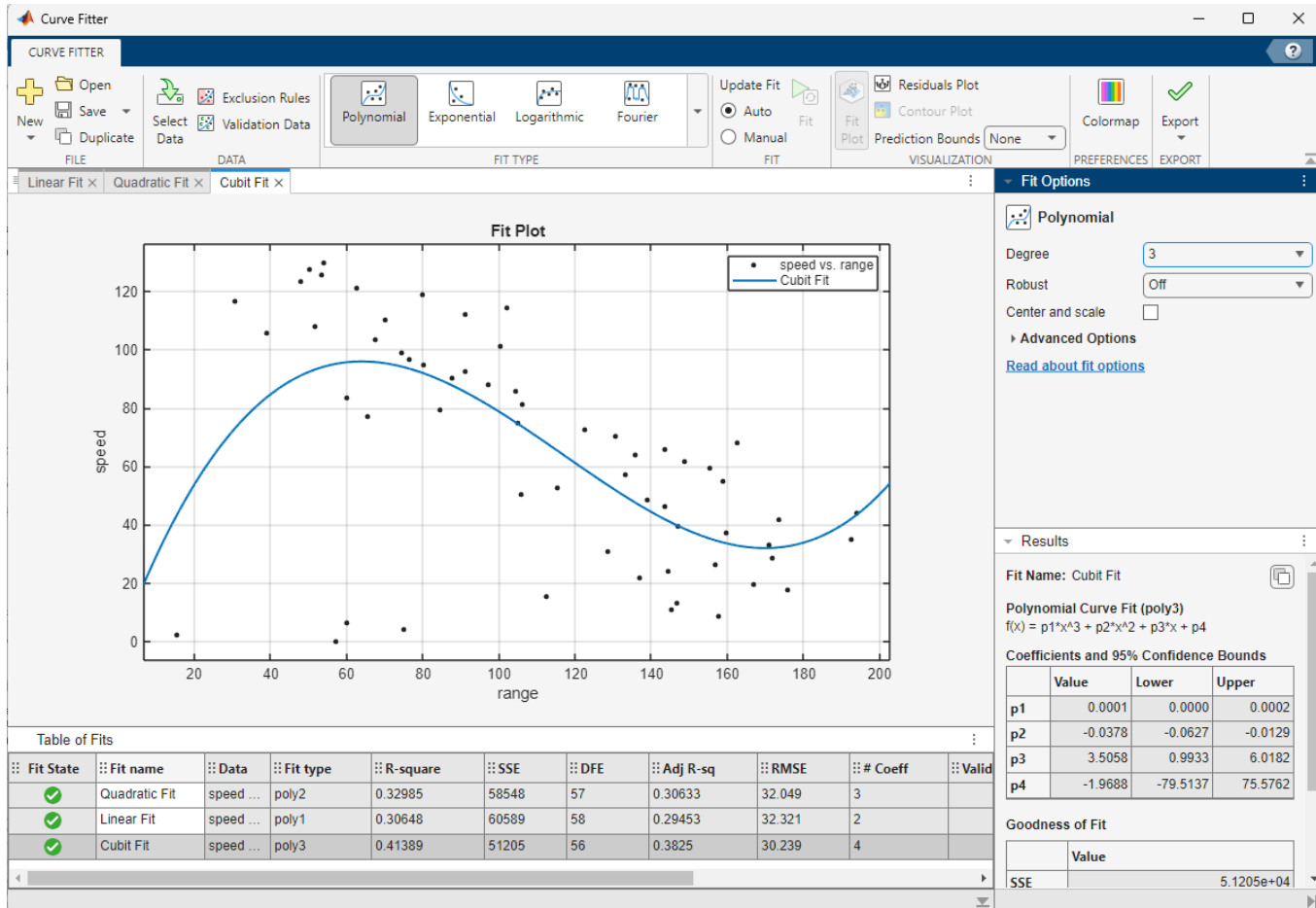
Curve Fitter App

- Let's try different curves and let's see to the "residuals"
- "Duplicate" at "Fit State" and create two Windows for Linear and Quadratic fit



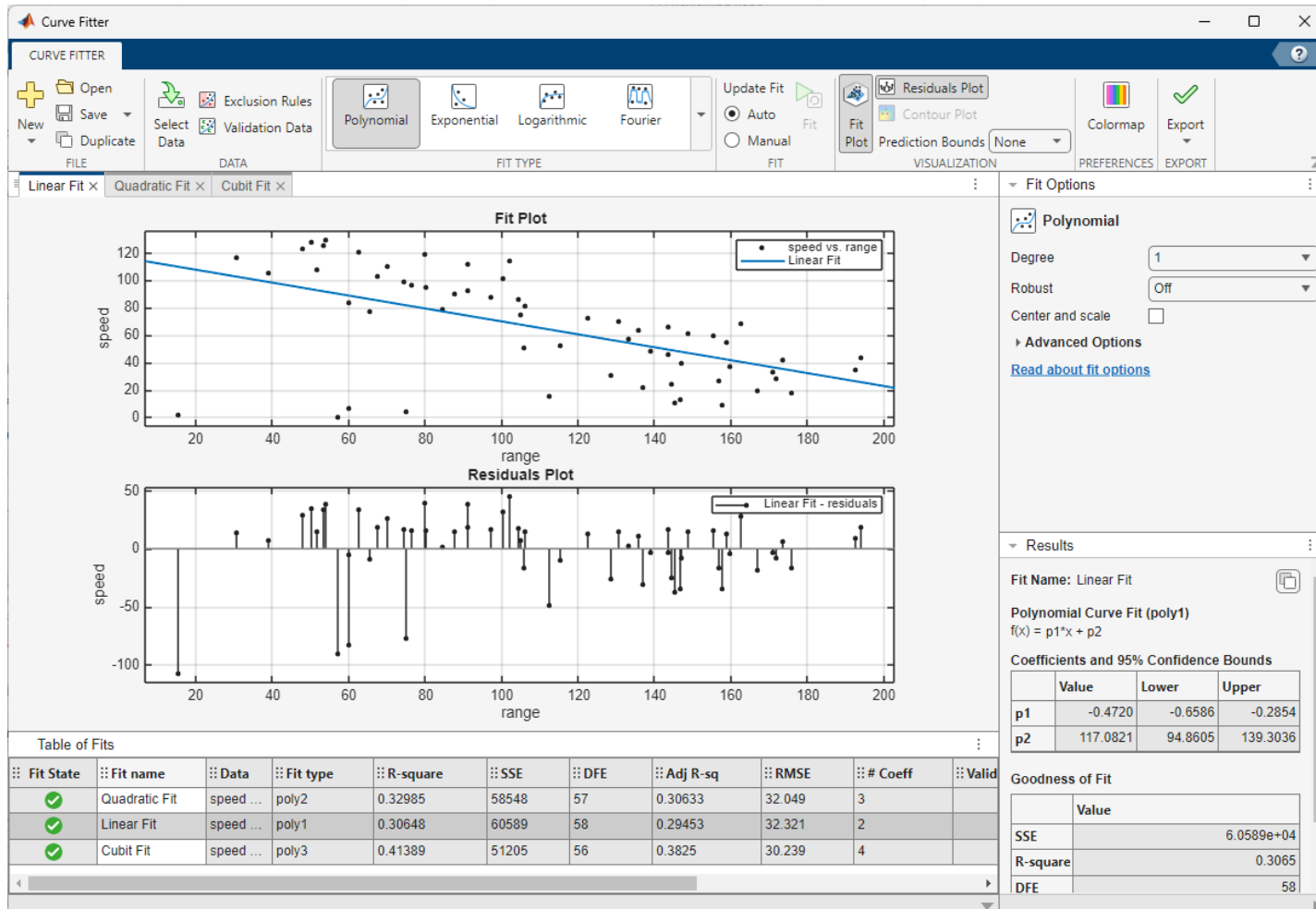
Curve Fitter App

- And “Cubic” one



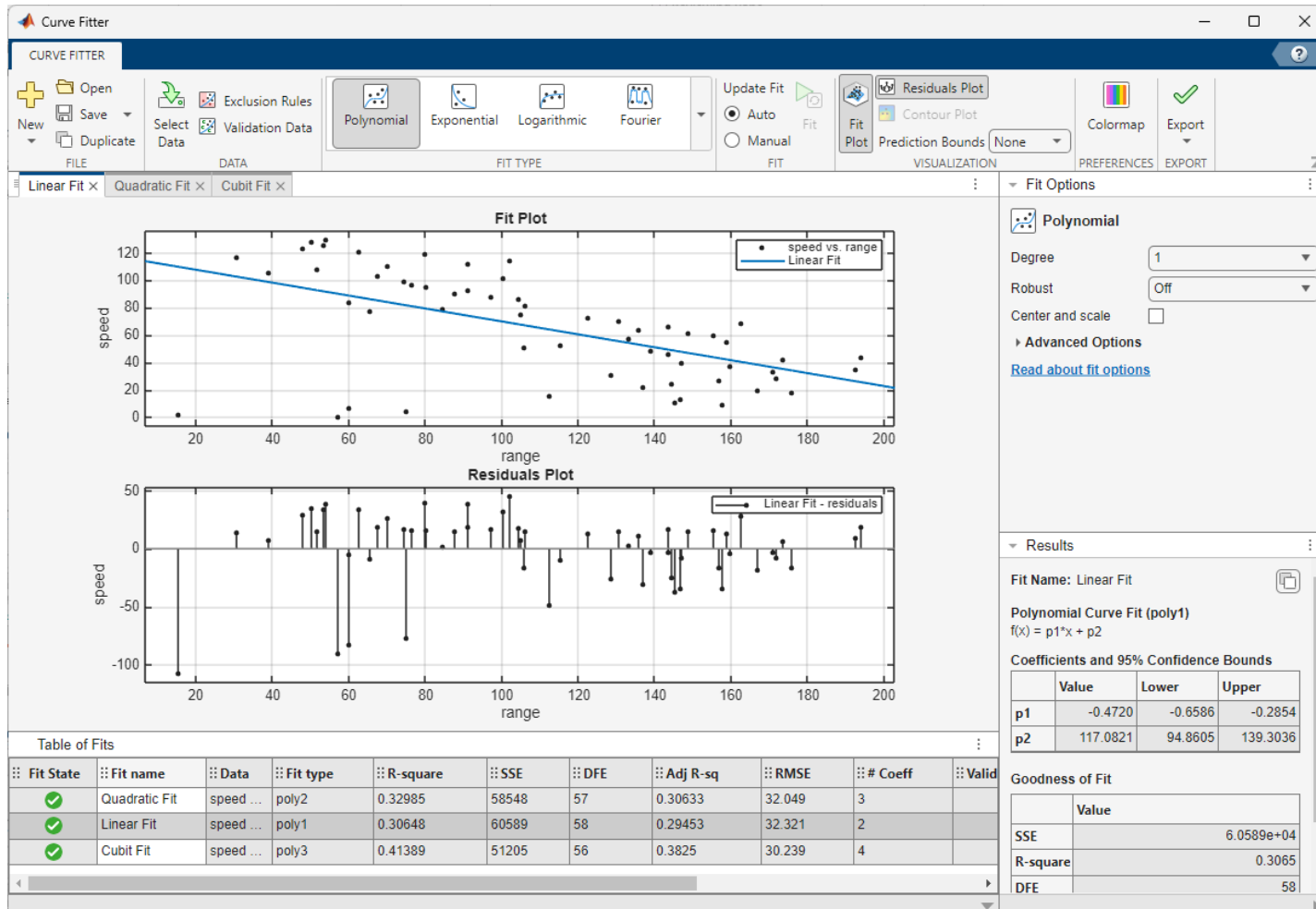
Curve Fitter App

- Now we can see the residuals for “Linear Fit”



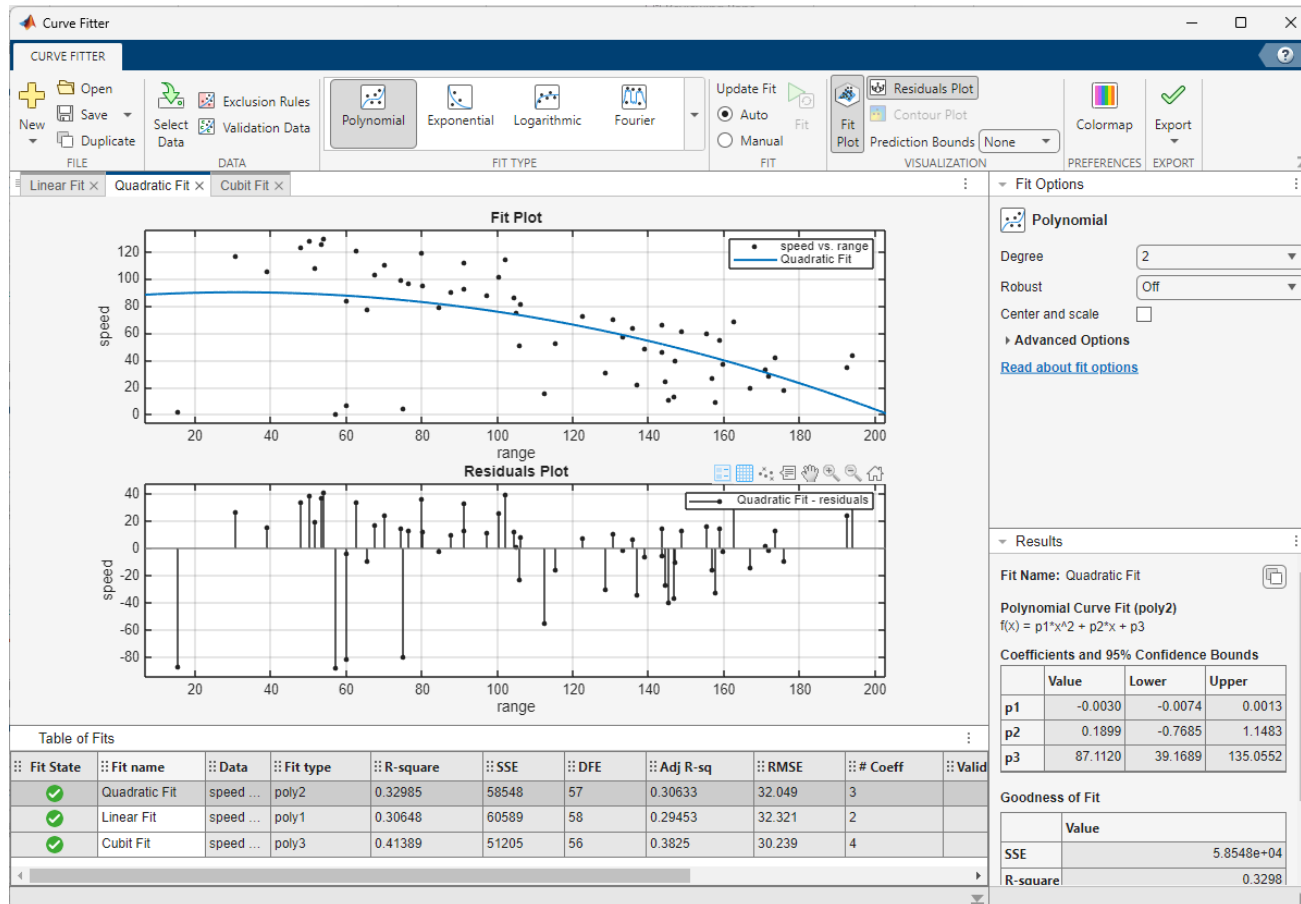
Curve Fitter App

- Now we can see the residuals for “Linear Fit”



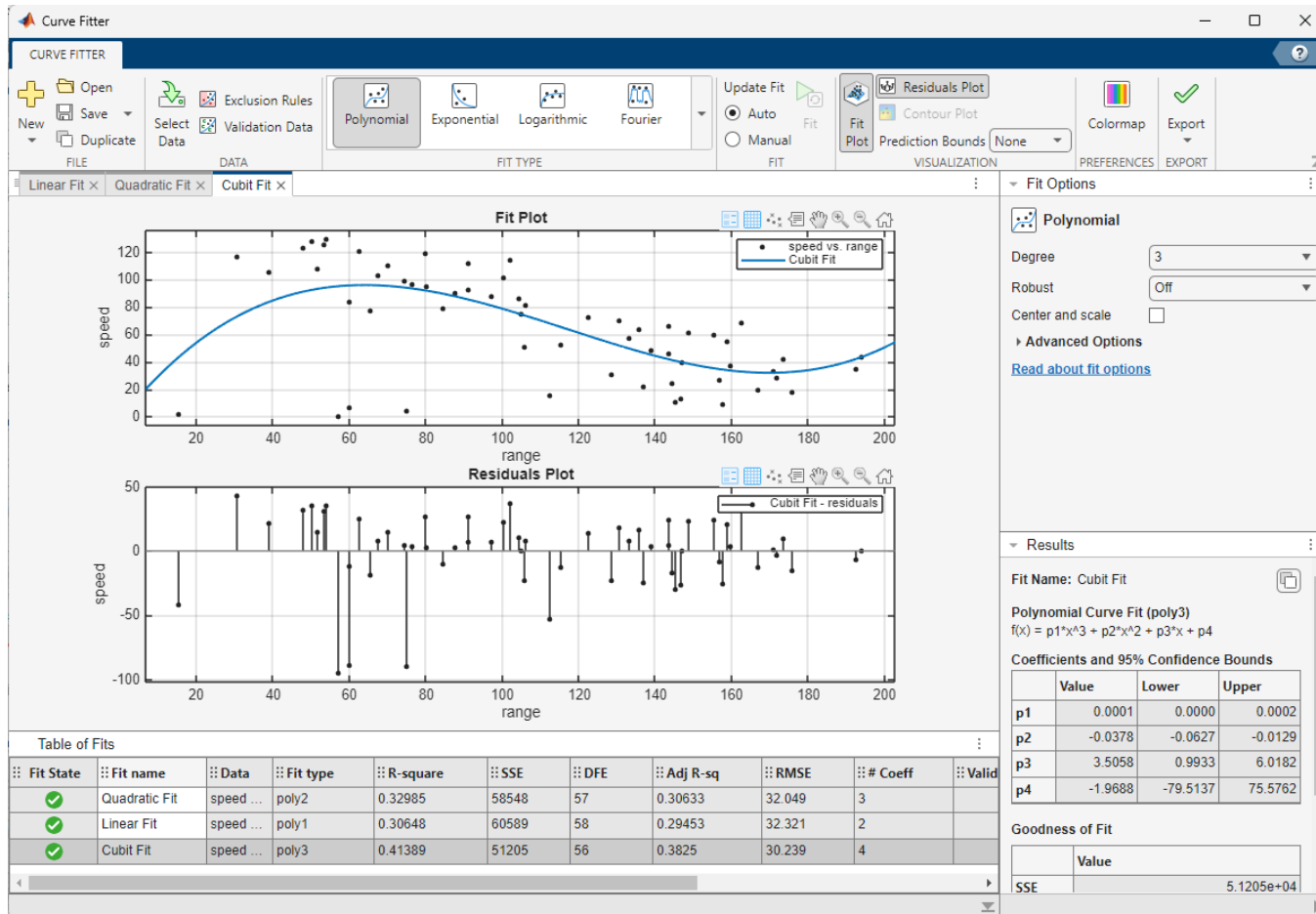
Curve Fitter App

- And the residuals for “Quadratic Fit”



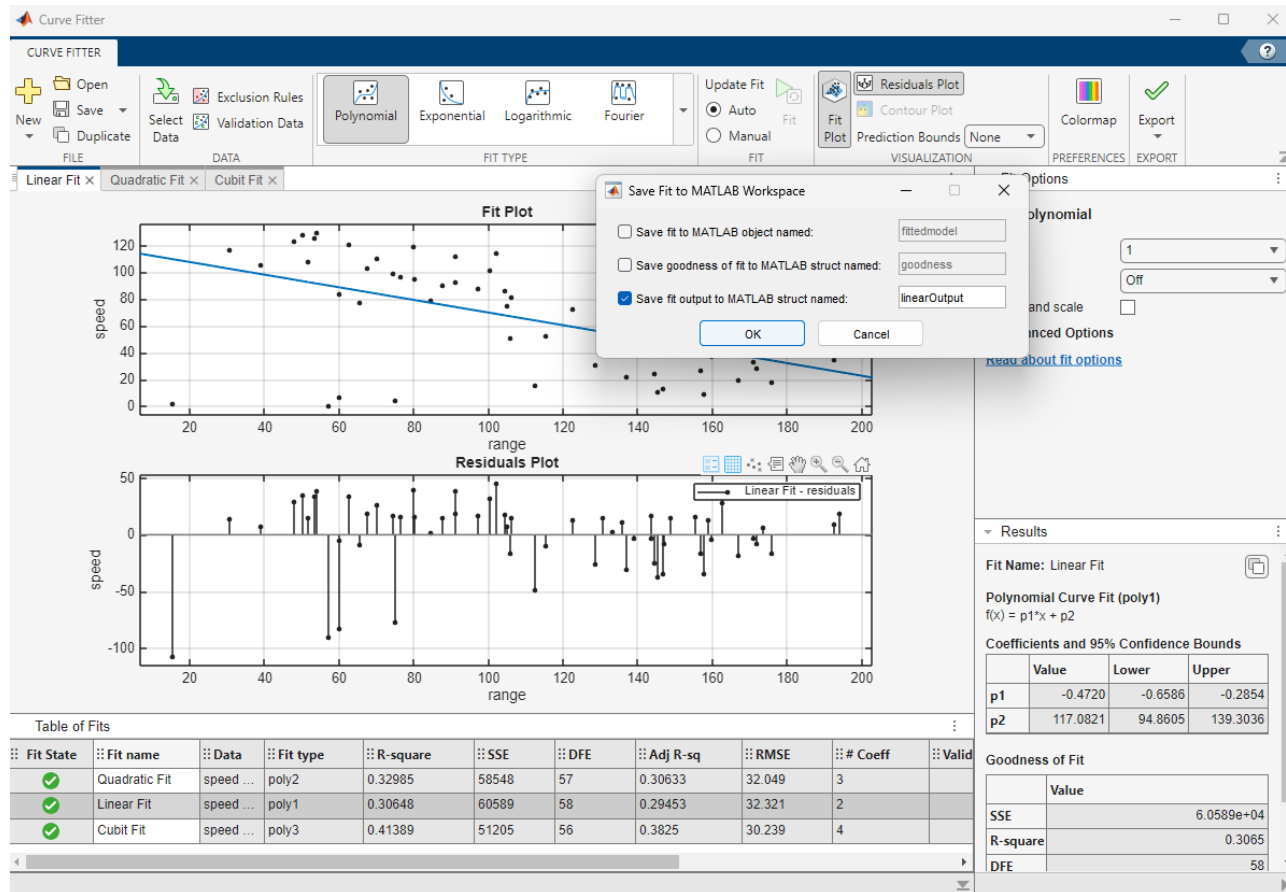
Curve Fitter App

- And for “Cubic Fit”



Curve Fitter App

- You can also “save” the Linear results to your Matlab “Workspace” as “linearOutput”.



Curve Fitter App

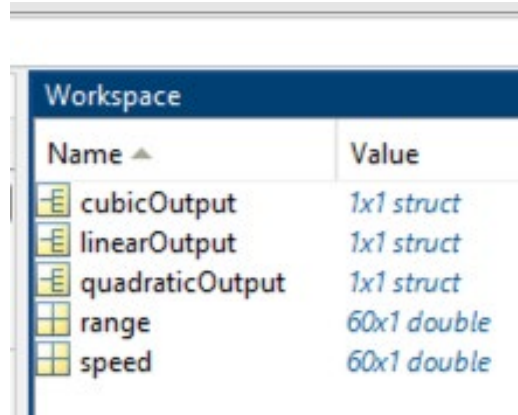
- You can also “save” the Linear results to your Matlab “Workspace” as “linearOutput”.
- Repeat “save” the “quadraticOutput”
- Same “save” the “cubicOutput”
- Save the 3 fits in a file using:
“save fitOutput.mat cubicOutput linearOutput quadraticOutput”
- Now you should have:

Current Folder	
Name ▲	Date Modified
.MATLABDriveTag	2/2/2025 9:47 PM
CompareMultFits2.sfit	1/31/2025 6:12 PM
evData.mat	1/31/2025 5:01 PM
fitOutput.mat	2/2/2025 10:13 PM
multipleEVData.mat	1/31/2025 5:01 PM
TAIM_1.mlx	2/2/2025 10:05 PM

fitOutput.mat (MAT-file)	
Name	Value
cubicOutput	1x1 struct
linearOutput	1x1 struct
quadraticOutput	1x1 struct

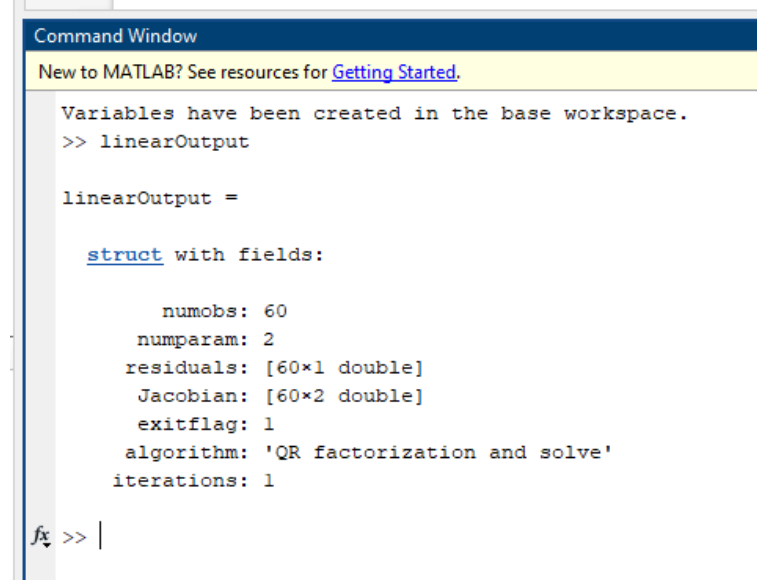
Curve Fitter App

- You can also “save” the Linear results to your Matlab “Workspace”



Name	Value
cubicOutput	1x1 struct
linearOutput	1x1 struct
quadraticOutput	1x1 struct
range	60x1 double
speed	60x1 double

- You can use the “Command” Window.



```

Command Window
New to MATLAB? See resources for Getting Started.

Variables have been created in the base workspace.
>> linearOutput

linearOutput =

    struct with fields:

        numobs: 60
        numparam: 2
        residuals: [60x1 double]
        Jacobian: [60x2 double]
        exitflag: 1
        algorithm: 'QR factorization and solve'
        iterations: 1

fx >> |
  
```

Curve Fitter App

- Output is organized as a Matlab structure variable

Live Editor - TAIM_1.mlx *

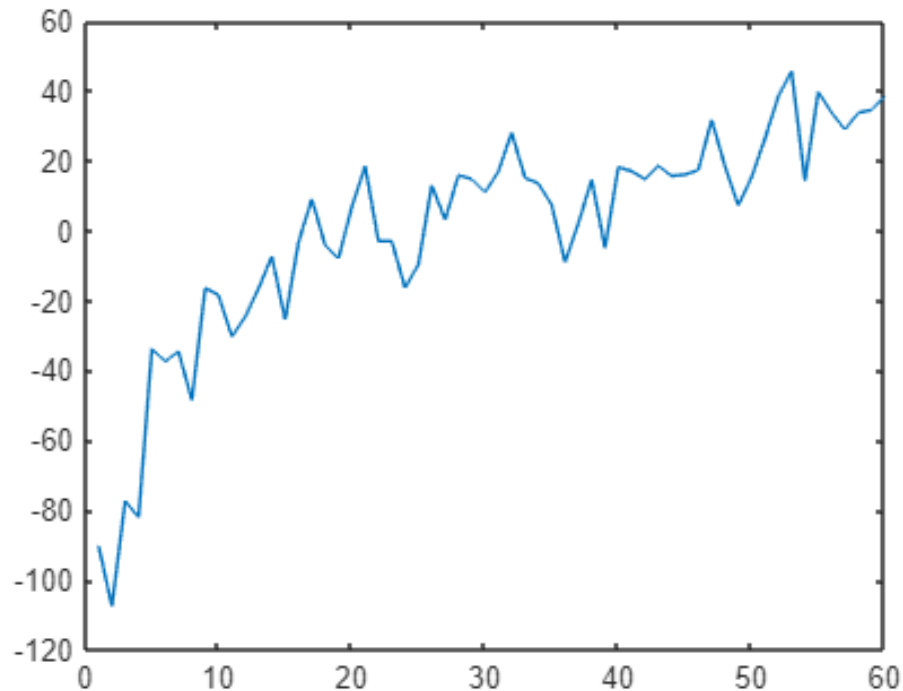
linearOutput ✕

1x1 struct with 7 fields

Field ▲	Value
numobs	60
numparam	4
residuals	60x1 double
Jacobian	60x4 double
exitflag	1
algorithm	'QR factorization and...
iterations	1

Curve Fitter App

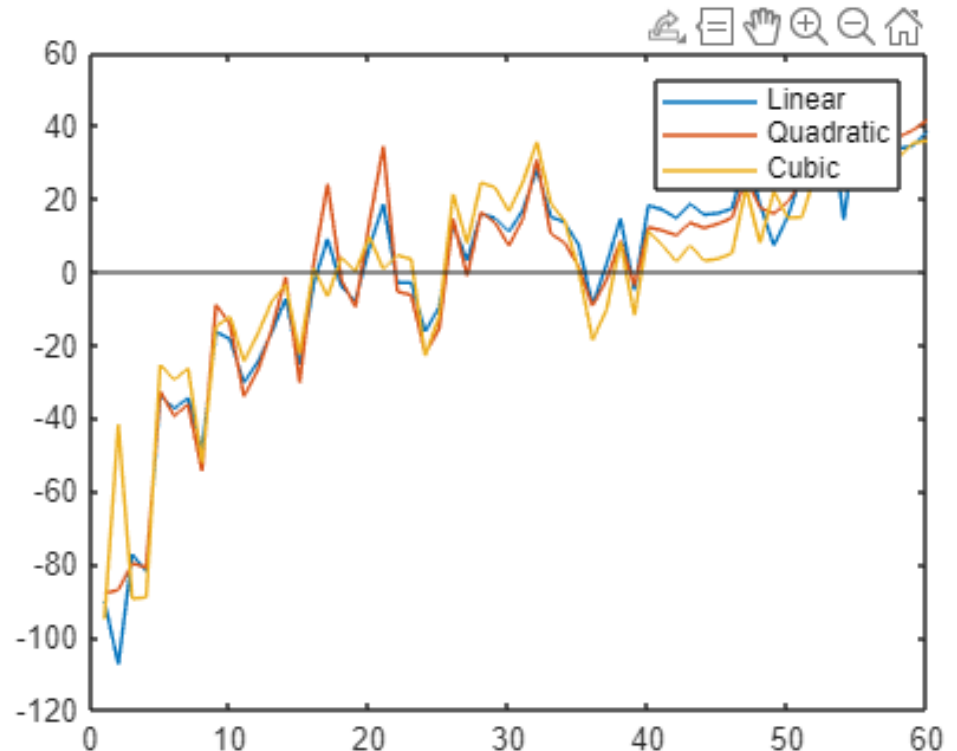
- Access to the residuals of the linear fit and make a plot of the residuals...
- You should get:



Curve Fitter App

- Using
hold on
plot(x)
hold off

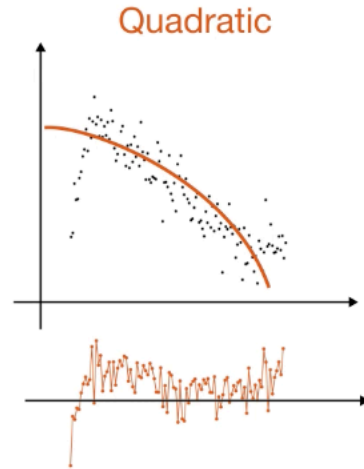
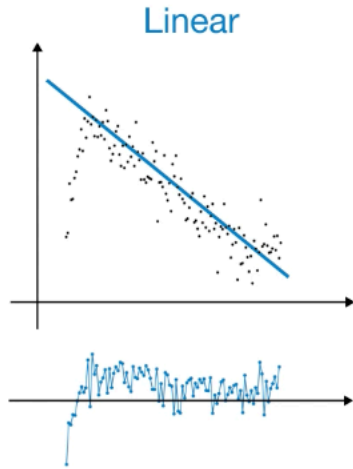
- ➔ Add plots of the residuals for the quadratic and cubic fits to the existing plot. Add a legend that labels the plots "Linear", "Quadratic", and "Cubic".
- ➔ You can also add a line at "0": "yline(0)"



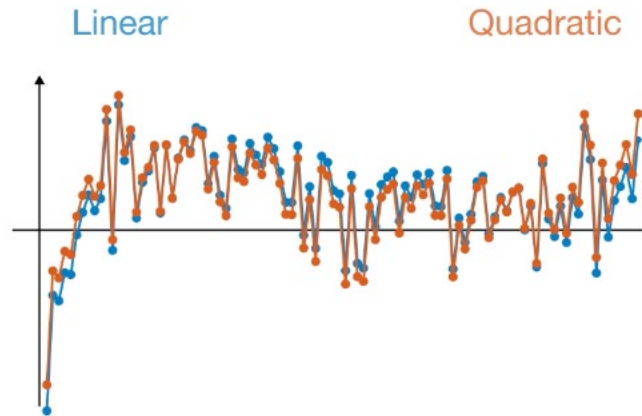
- You should get:

Notice that the fits improve as the residuals move closer and closer to zero. However, this method for comparing different models remains subjective.

Which is the best fit?

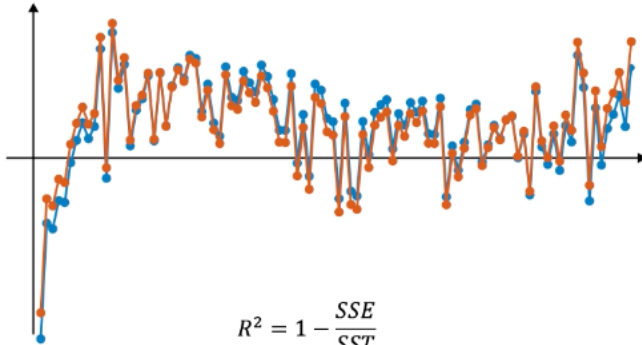


Notice that the fits improve as the residuals move closer and closer to zero. However, this method for comparing different models remains subjective.



Which is the best fit?

$$SSE = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$



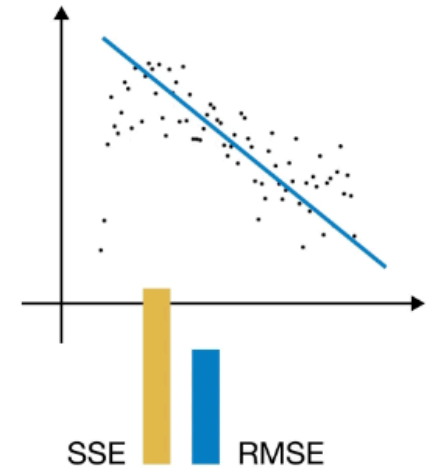
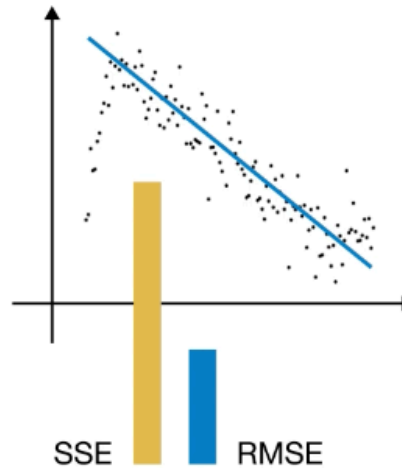
$$R^2 = 1 - \frac{SSE}{SST}$$

Mathematical formulas to quantitatively evaluate the goodness of fit.

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}}$$

Sum Squares Error (SSE) does not count on that 2 curves may have different number of points.

Root Mean Square Error (RMSE) is better.



Which is the best fit?

All this data
automatically
calculated by the App

Results			
Fit Name: Cubic Fit			
Polynomial Curve Fit (poly3)			
$f(x) = p1 \cdot x^3 + p2 \cdot x^2 + p3 \cdot x + p4$			
Coefficients and 95% Confidence Bounds			
	Value	Lower	Upper
p1	0.0001	0.0000	0.0002
p2	-0.0378	-0.0627	-0.0129
p3	3.5058	0.9933	6.0182
p4	-1.9688	-79.5137	75.5762
Goodness of Fit			
	Value		
SSE	5.1205e+04		
R-square	0.4139		
DFE	56		
Adj R-sq	0.3825		
RMSE	30.2386		

Which is the best fit?

All this data automatically calculated by the App

Fit State	Fit name	Data	Fit type	R-square	SSE	DFE	Adj R-sq	RMSE	# Coeff	Validation Data	Validation SSE	Validation RMSE
✔	Linear Fit	r vs. v	poly1	0.59166	2.4579e+05	42	0.58194	76.499	2			
✔	Quadratic Fit	r vs. v	poly2	0.65864	2.0547e+05	41	0.64198	70.792	3			
✔	Cubic Fit	r vs. v	poly3	0.75223	1.4913e+05	40	0.73365	61.06	4			

$$R^2 = 1 - \frac{SSE}{SST}$$

$$SSE = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}}$$

Which is the best fit?

Task:


- 1 Generate a fourth-order (quartic) fit and a fifth-order (quintic) fit.
- 2 Export the goodness-of-fit statistics for the quartic fit to a variable named "goodness4".

See: SSE and R-square of Table of Fits.

- SSE decreases while R-square increases -> indicating that the fits keep improving.
- Adding more terms to the polynomial results in better fits.
- So, should you continue fitting more complicated models with more and more terms?
- It might start fitting random noise in the data instead of just the general trend.
- This result is called **overfitting**.

Which is the best fit?

- This result is called **overfitting**.
- One easy way to see if you're overfitting the data is to look at the best fit parameter values.
- Notice that the two lowest order parameters in the fifth-order model are essentially zero. This result can be an indicator that you're overfitting the data.

Results			
Fit Name: Quintic Fit  Equation is badly conditioned. Try centering and scaling, or add points at non-repeated x values.			
Polynomial Curve Fit (poly5) $f(x) = p1 \cdot x^5 + p2 \cdot x^4 + p3 \cdot x^3 + p4 \cdot x^2 + p5 \cdot x + p6$			
Coefficients and 95% Confidence Bounds			
	Value	Lower	Upper
p1	0.0000	-0.0000	0.0000
p2	-0.0000	-0.0000	0.0000
p3	0.0025	-0.0013	0.0062
p4	-0.2572	-0.5921	0.0777
p5	12.3200	-0.9846	25.6245
p6	-116.2946	-301.1980	68.6089
Goodness of Fit			
	Value		
SSE	4.9494e+04		
R-square	0.4335		
DFE	54		
Adj R-sq	0.3810		
RMSE	30.2747		

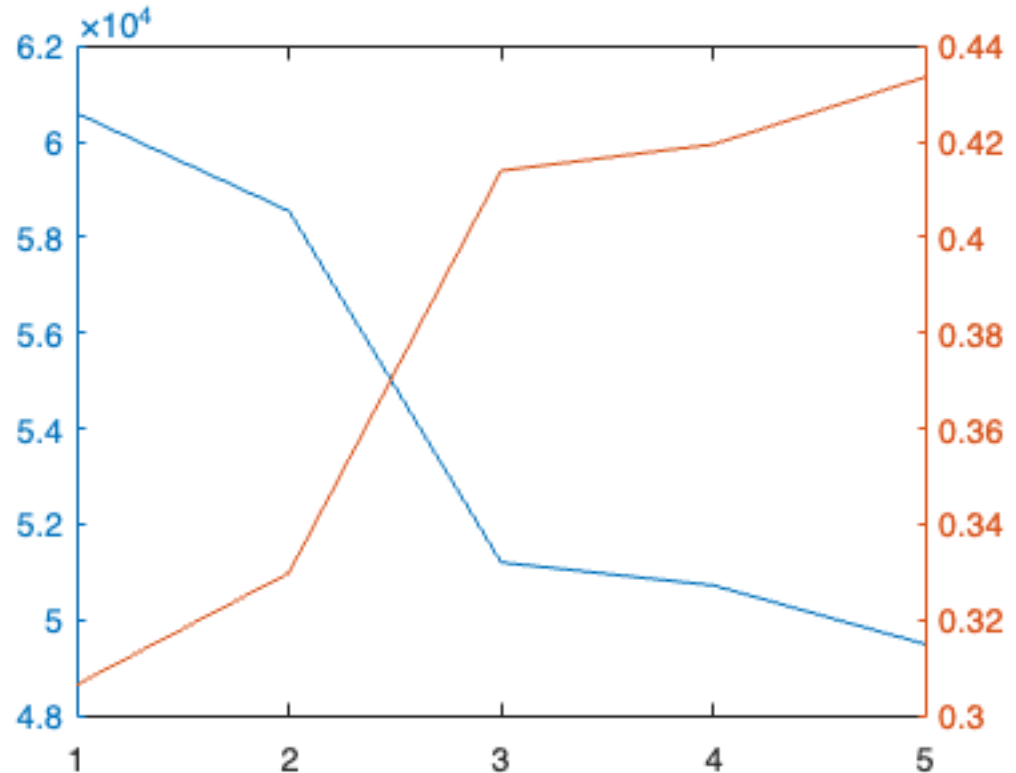
Which is the best fit?

- Once you have saved all the "goodness1" to "goodness5"
- Generate an overall "goodness" array by:
"A = [array1 array2 array3]"

("goodness = [goodness1 goodness2 goodness3 goodness4 goodness5]")
- Typically, handling data is easier if you use tables instead of structures, because you can access using just the names of the columns.
- Convert a struct variable to a table by: "table = struct2table(structure)"

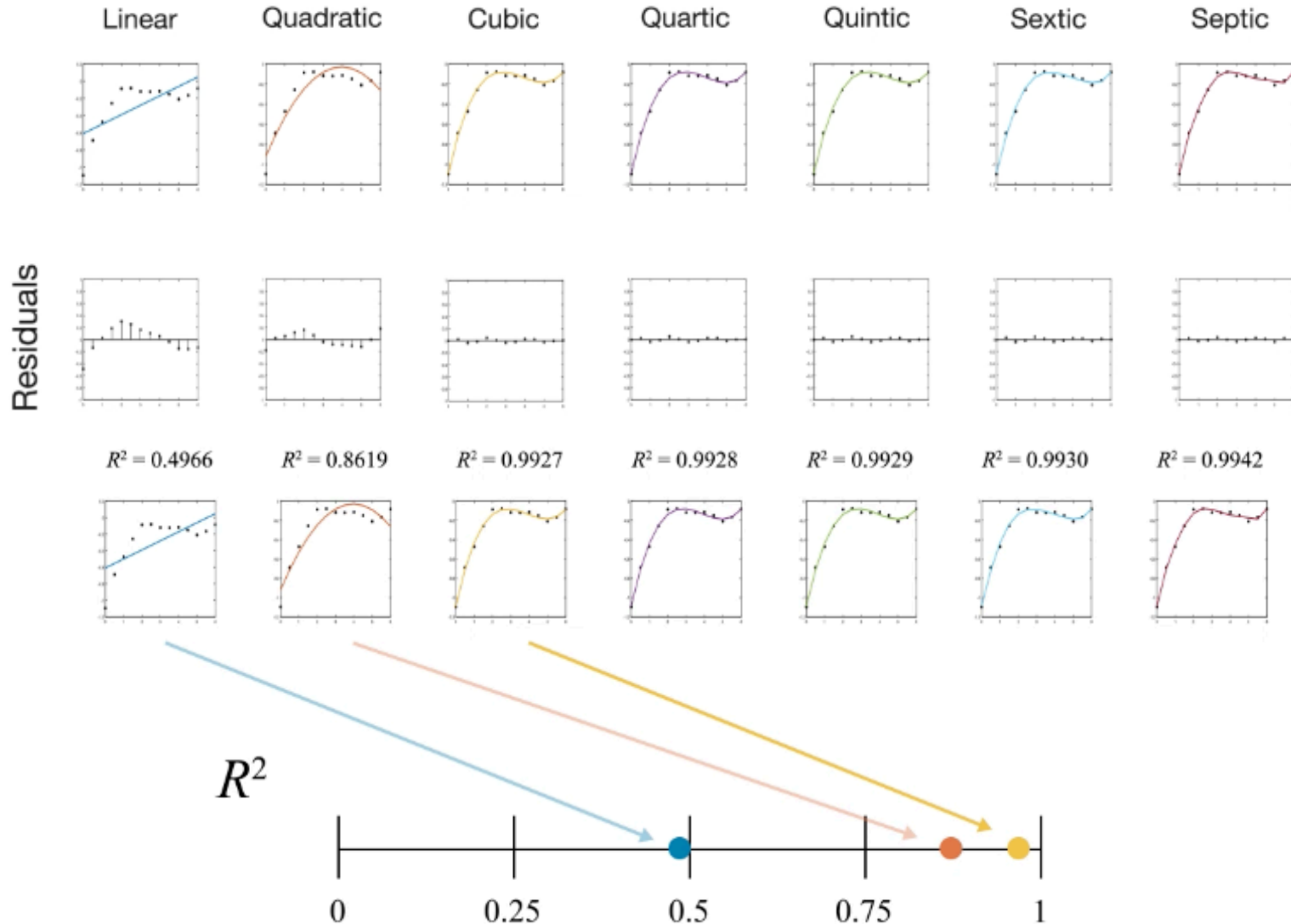
Which is the best fit?

- Using:
yyaxis command.
yyaxis left
plot(x)
yyaxis right
plot(y)
- Plot the SSE and R-squared values on the same figure using the yyaxis command. Plot the SSE on the left and the R-squared on the right.

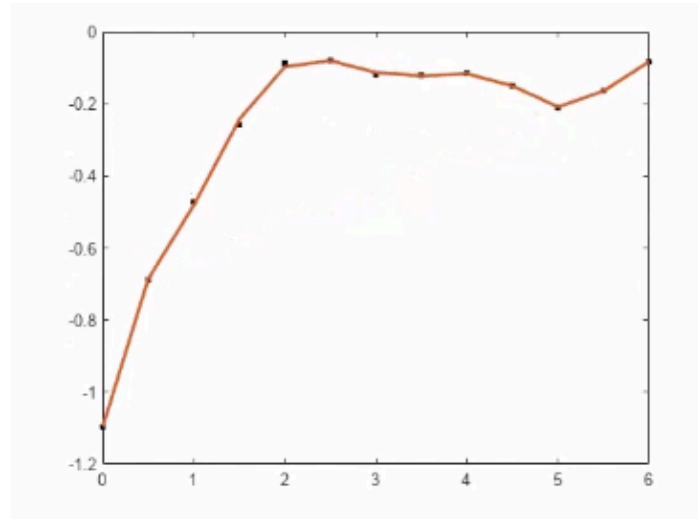


Both the SSE and R-squared look opposite. However, both seem to follow the same pattern. The goodness-of-fit improves significantly until the cubic model and then begins to plateau. The quintic model has the best SSE and R-squared.

Which is the best fit?



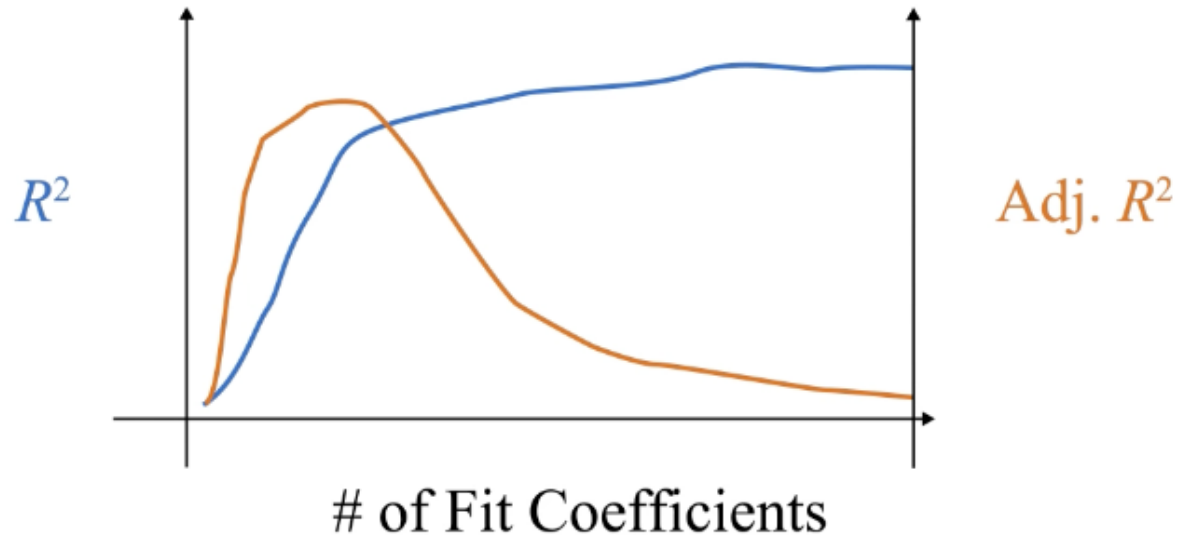
Which is the best fit?



$$y = c_1 + c_2x + c_3x^2 + c_4x^3 + c_5x^4 + c_6x^5 + c_7x^6 + c_8x^7 + c_9x^8 + c_{10}x^9 + c_{11}x^{10} + c_{12}x^{11} + c_{13}x^{12}$$

- Just with the R-squared values, you might be fitting not only the underlying trend, but also the noise in the data.
- This is called **overfitting the data**,
- It can be avoided by using a modified version of the R-squared called the **adjusted R-squared**.

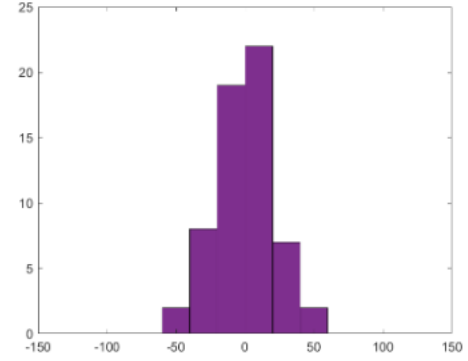
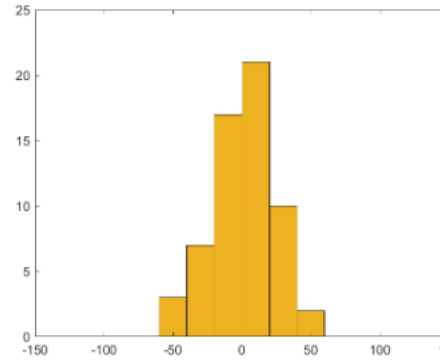
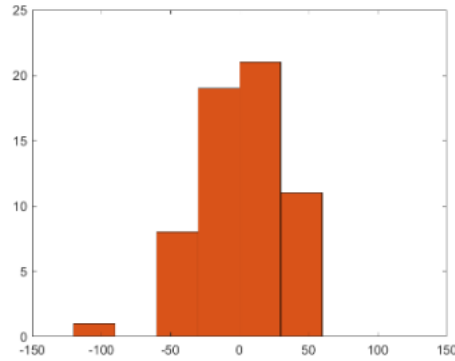
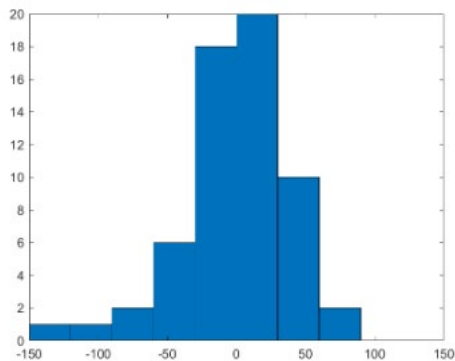
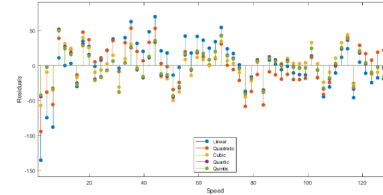
Which is the best fit?



- R-squared increases with the addition of more fit coefficients
- Adjusted R-squared only increases if the more complicated model results in a sufficiently better fit
- The idea is to use a simpler model if possible, or a more complex one that justifies the use of an additional coefficient.

Hints to select the best fit

1. Visual inspection of the fits



Linear — and quadratic Residuals
Some significant outliers @ -150.
-> to conclude that these models are not ideal
for this particular data set.

Cubic Residuals
No significant
outliers -> better
fitting.

Quartic Residuals
The residuals are
only marginally
better compared
to cubic model.

Hints to select the best fit

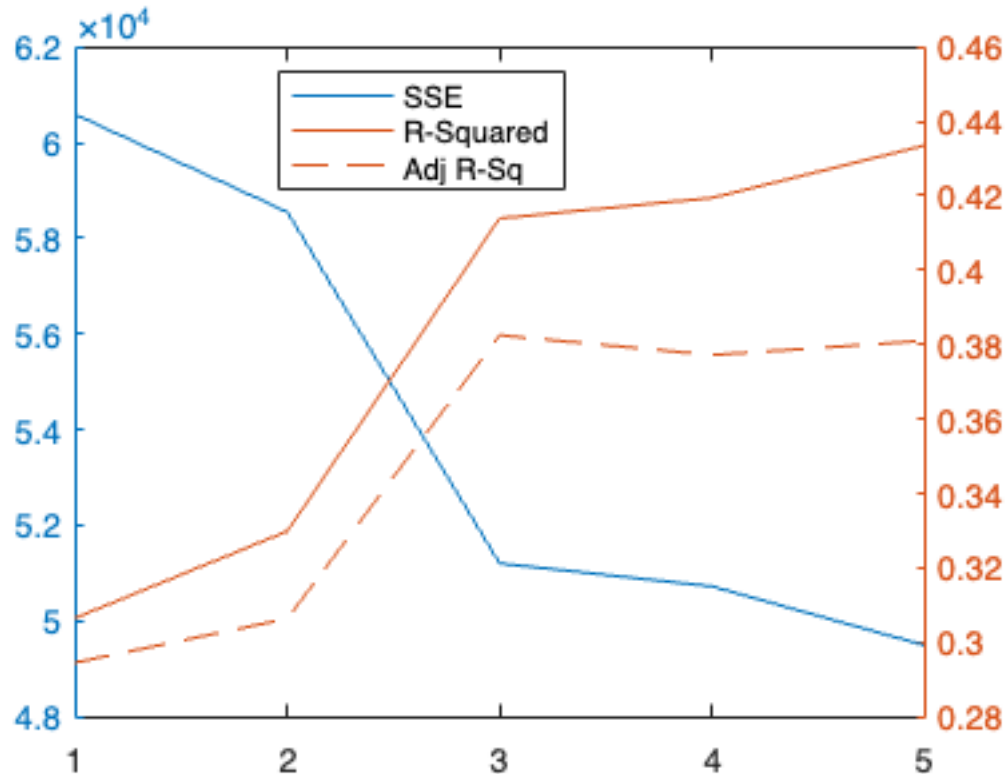
2. Look at the Goodness-of-Fit Statistics

Table of Fits									
Fit State	Fit name	Data	Fit type	R-square	SSE	DFE	Adj R-sq	RMSE	# Coeff
✓	Quadratic Fit	speed ...	poly2	0.32985	58548	57	0.30633	32.049	3
✓	Linear Fit	speed ...	poly1	0.30648	60589	58	0.29453	32.321	2
✓	Cubic Fit	speed ...	poly3	0.41389	51205	56	0.3825	30.239	4
✓	Quartic Fit	speed ...	poly4	0.4194	50723	55	0.37718	30.368	5
⚠	Quintic Fit	speed ...	poly5	0.43347	49494	54	0.38102	30.275	6

- SSE and R-squared almost always increase continually for models with an increasing number of fitting coefficients.
- How much they improve? -> This can indicate if starting to overfit data

Hints to select the best fit

2. Look at the Goodness-of-Fit Statistics



- Adjusted R^2 does not continually increase because it penalizes models for their number of fit coefficients.

Hints to select the best fit

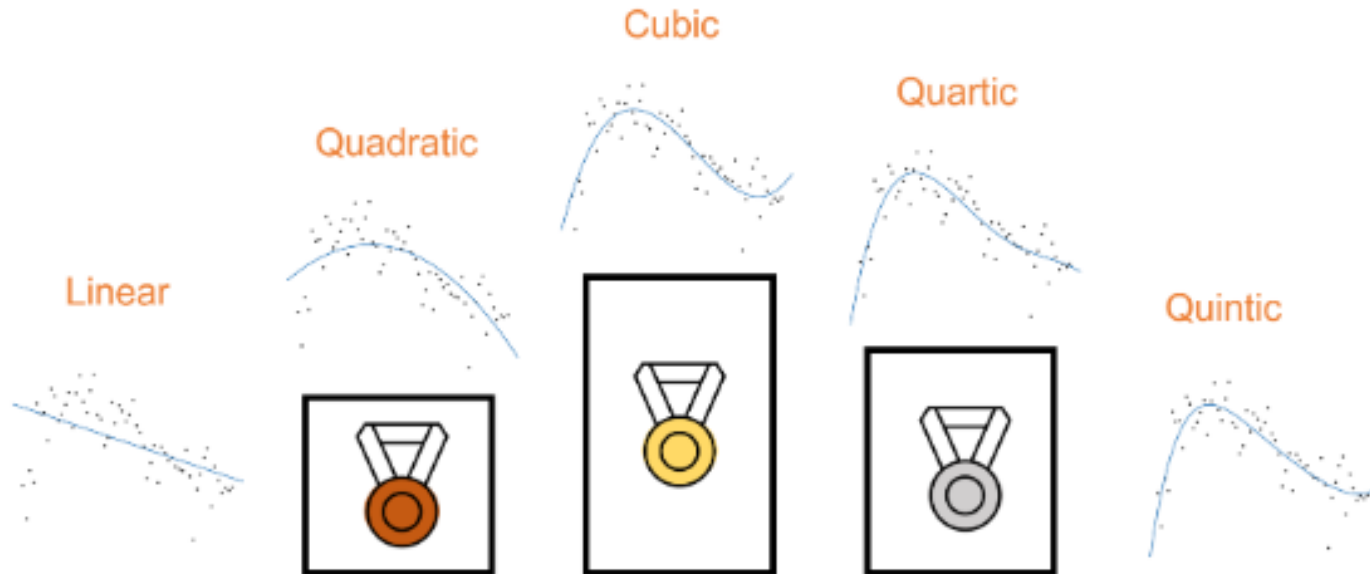
3. Look at the Best Fit Coefficients and Confidence Bounds

If a model contains any unnecessary fit coefficients, -> the best fit values of those coefficients will likely be very close to zero (if the 95% confidence bounds contain a negative lower bound and a positive upper bound).

(The confidence bounds indicate that the best fit value for a fit coefficient has a 95% chance of lying inside that range.)

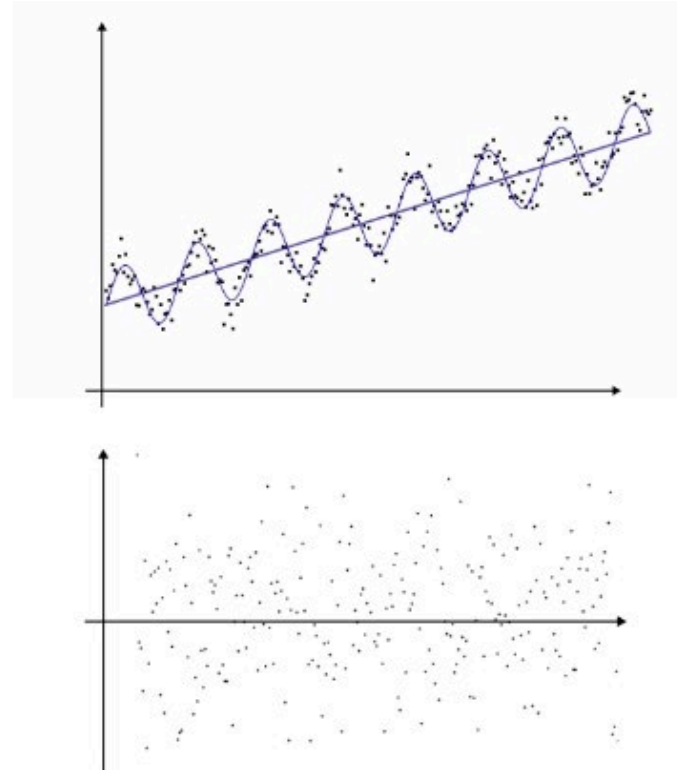
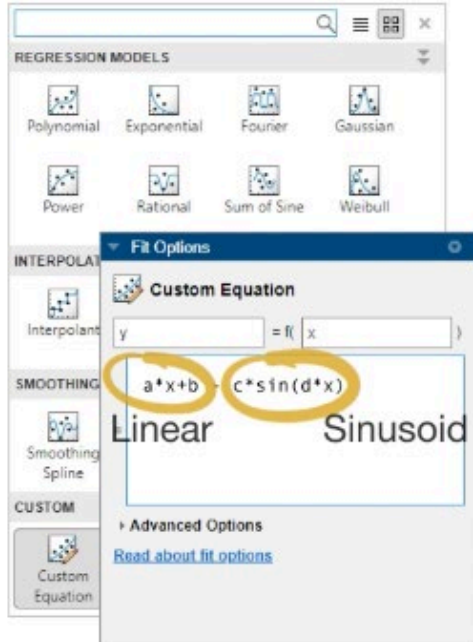
Results			
Fit Name: Cubit Fit			
Polynomial Curve Fit (poly3)			
$f(x) = p1 \cdot x^3 + p2 \cdot x^2 + p3 \cdot x + p4$			
Coefficients and 95% Confidence Bounds			
	Value	Lower	Upper
p1	0.0001	0.0000	0.0002
p2	-0.0378	-0.0627	-0.0129
p3	3.5058	0.9933	6.0182
p4	-1.9688	-79.5137	75.5762
Goodness of Fit			
	Value		
SSE	5.1205e+04		
R-square	0.4139		
DFE	56		
Adj R-sq	0.3825		
RMSE	30.2386		

Which model do you rank as the best?



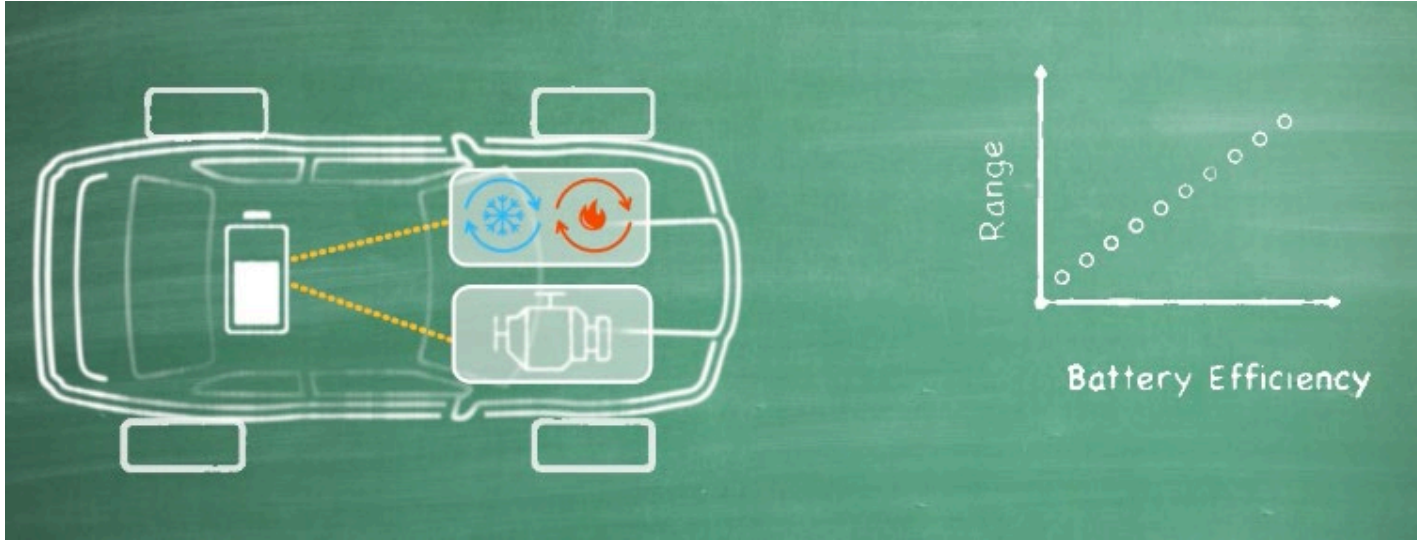
- The cubic model is the first to have best fit coefficients that are not microscopically small.
- It has goodness-of-fit statistics that are significantly better than those of the quadratic model and very close to those of the more complicated quartic and quintic models.
- In doubt, choose the simpler model

Fitting your own custom model?



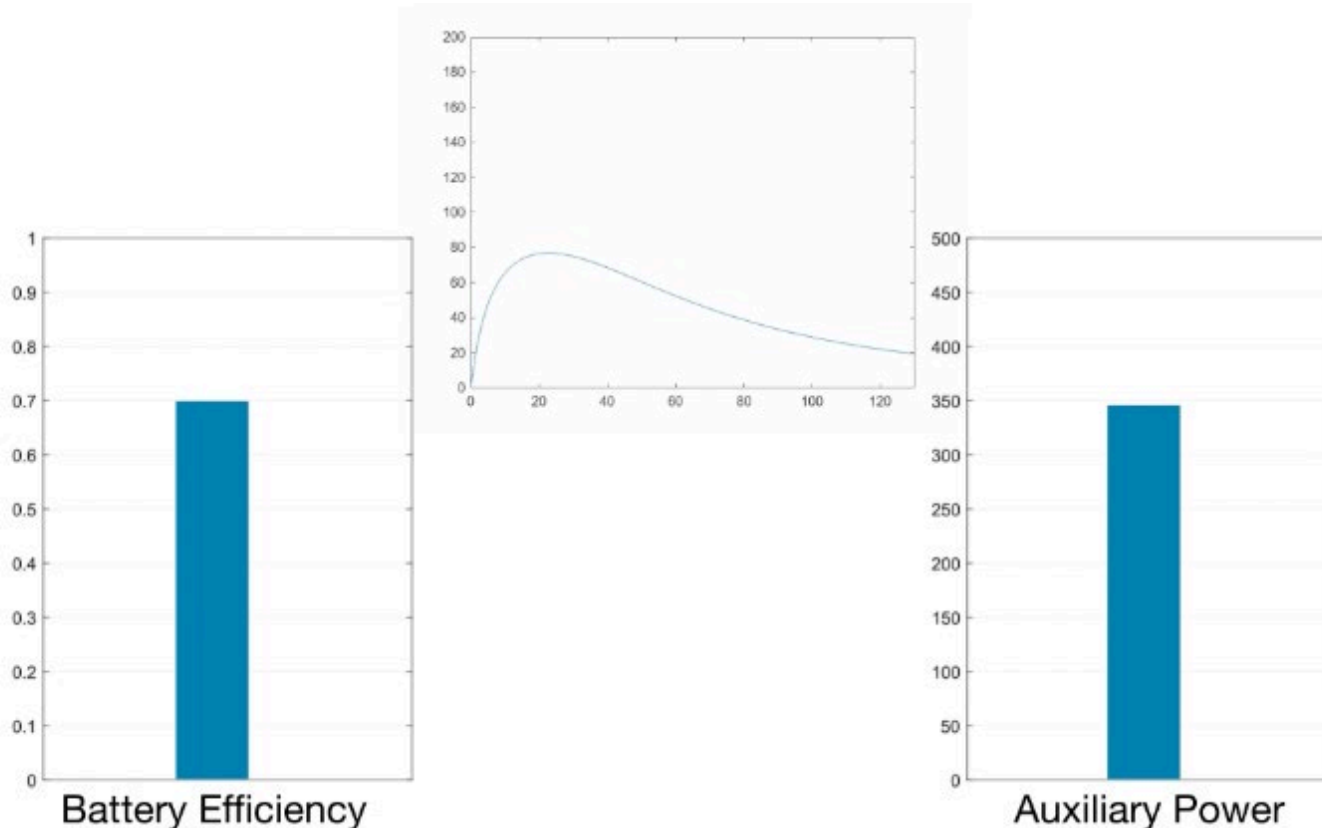
- Imagine you have a complex data that is not one of the common curves provided by Matlab.

Fitting your own custom model?



- Or you have a better physical model of the data, e.g. for an electrical vehicle, it is not only the consumption of the motor to move, but also,...

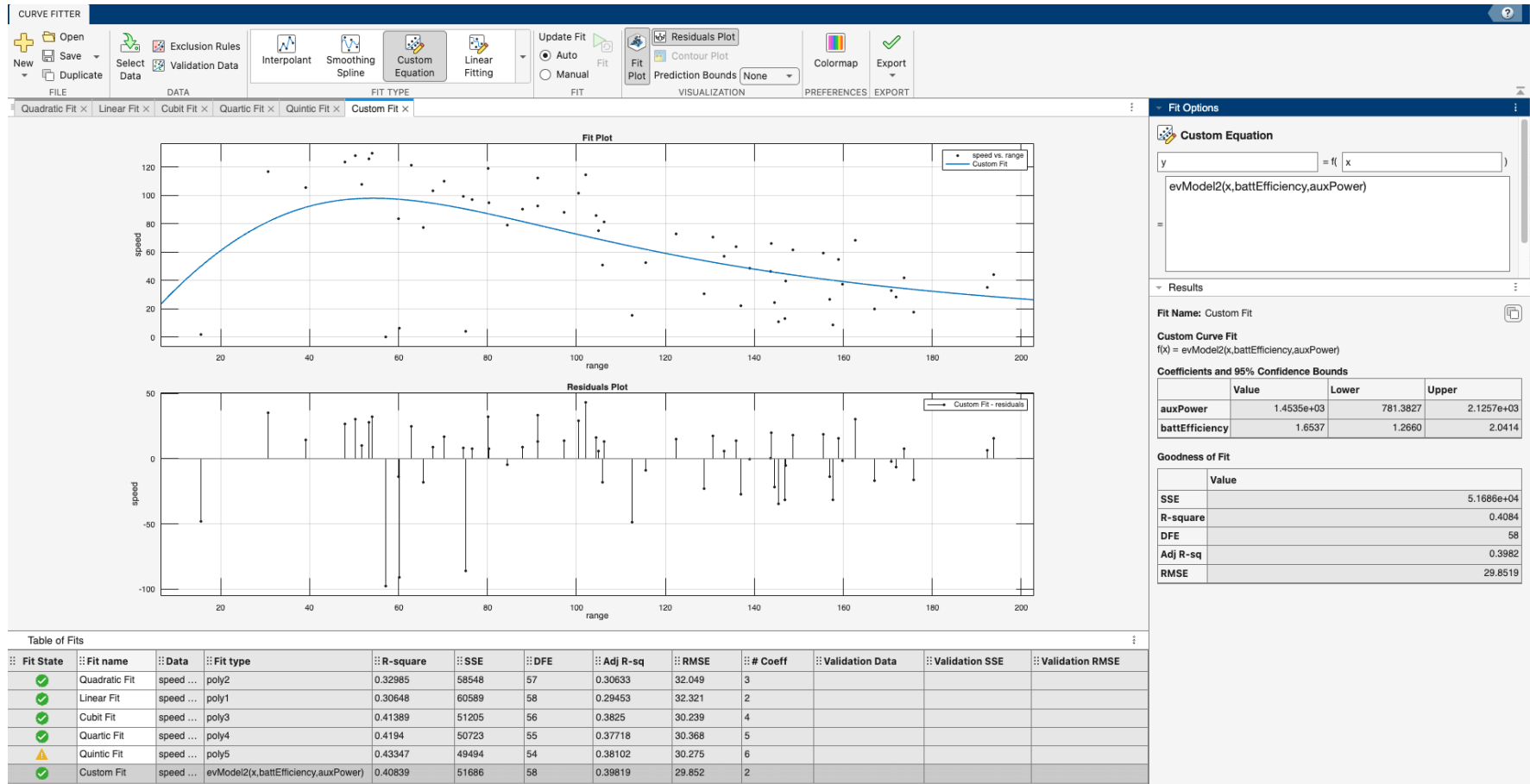
Fitting your own custom model?



- Or you have a better physical model of the data, e.g. for an electrical vehicle, it is not only the consumption of the motor to move, but also,... on the “Auxiliary power” consumption for air-conditioned, heating, music,...

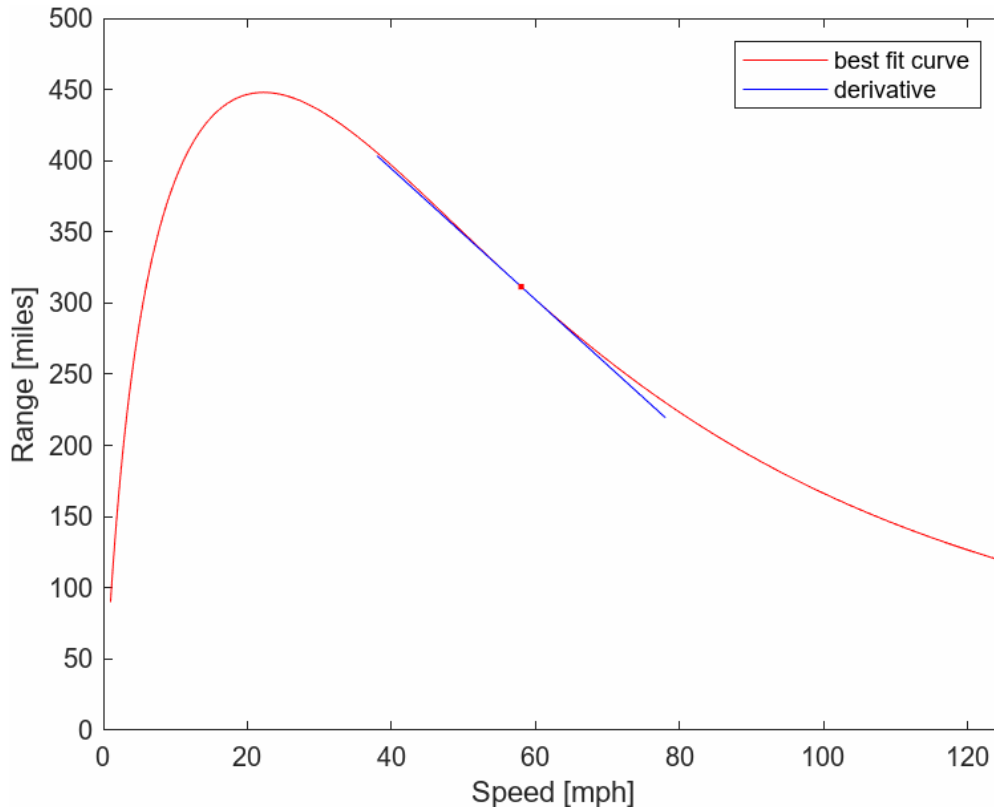
Fitting your own custom model?

- Follow the instructions of the Prof to build a function-based “interactive” model in your Matlab. -> You will be able to implement the custom fit, as:



Uses of a fit

- Now you can use the fit e.g. for finding the optimal speed:



Option 1: a) find the index of the maximum of the fitted model, b) index into the speed variable.

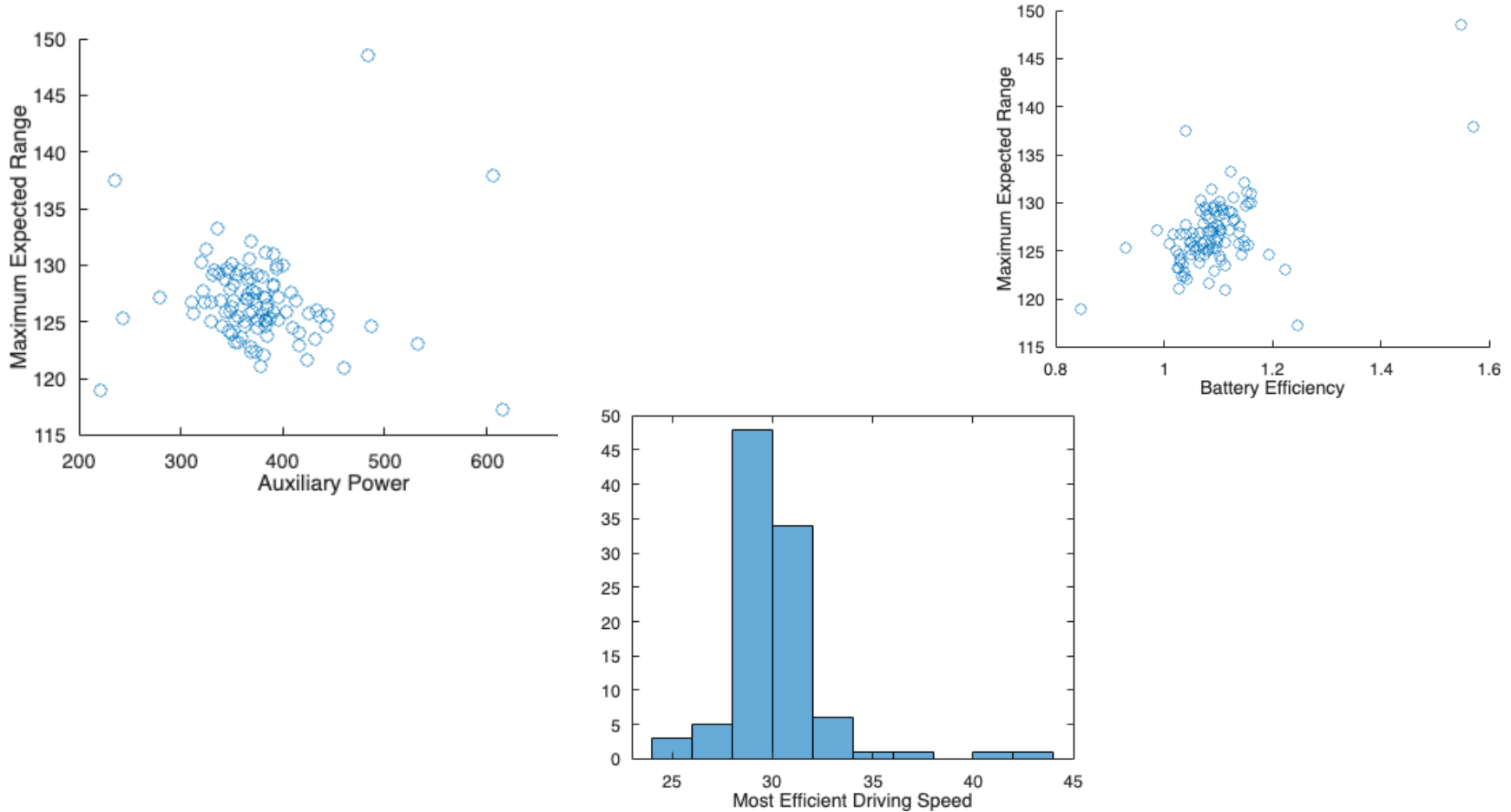
```
[maxVal,idx] = max(fittedmodel)
speedBest = speed(idx)
```

Option 2: Only a certain number of functions are compatible with fit objects.

Alternate approach ->
calculating slope = 0 -> find the corresponding speed.

Uses of a fit

- Follow the steps indicated by the Prof to automatize analysis to multiple vehicles



Now is your turn

- Now follow the next steps indicated by the Prof for your Task on this topic.
- Choose a data set that you want to analyse and develop your own analysis.
- You have also examples at:
 - https://es.mathworks.com/help/curvefit/getting-started-with-curve-fitting-toolbox.html?s_tid=CRUX_lftnav
 - https://es.mathworks.com/matlabcentral/fileexchange/93435-regression-basics?s_tid=ma_spoc_edu_orcf