Model Complexity, Overfitting, and Underfitting

Introduction

In this project, we first obtain a very complex model using “Linear regression with higher-order terms” with different values of N from one to ten as required. Then we reduce higher order terms (reduce complexity more) to get a polynomial with degree one also called an “Underfitted Model” or too simple. Then we use LASSO to remove the higher-order terms (reduce complexity) gradually to get a “Robust Model”.

Scatter Plotting the Dataset and the result obtained is shown below:

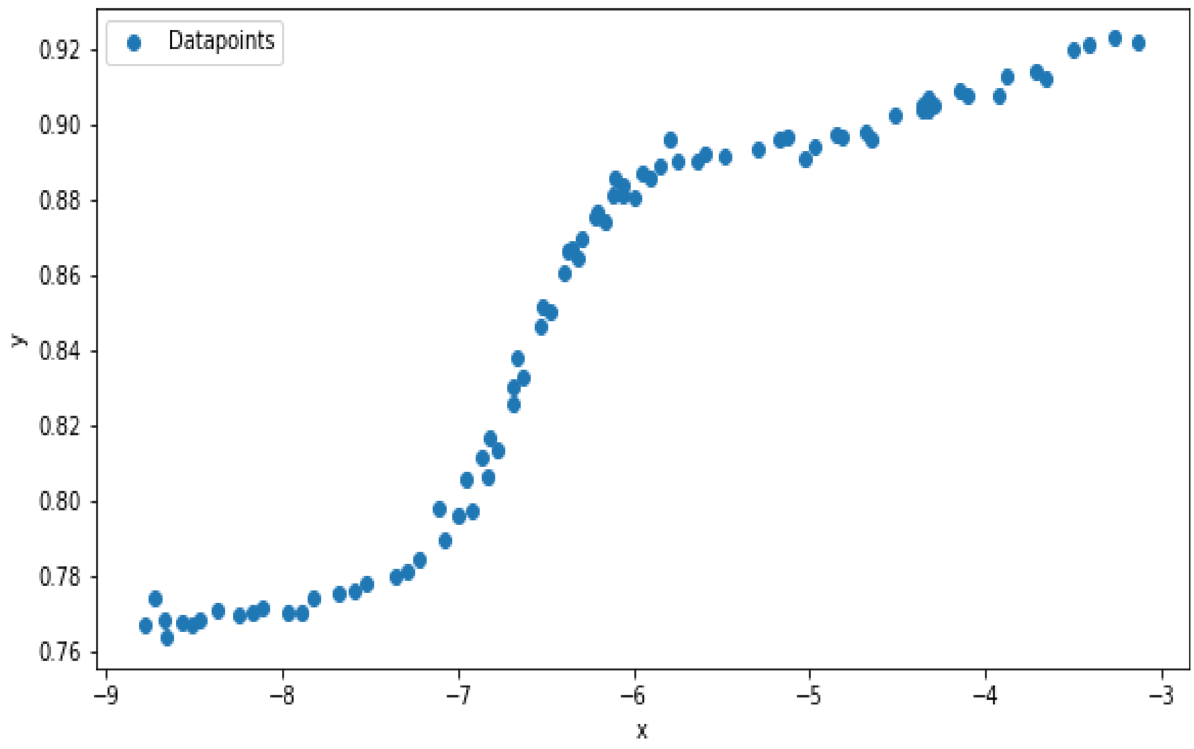


Fig 1: Scatter plot of the dataset without applying any model

Linear regression with higher-order terms

(a) What it is:

A linear regression model describes the relationship between a dependent variable, y, and one or more independent variables, X. The dependent variable is also called the response variable. Independent variables are also called explanatory or predictor variables. In linear regression, if a single independent variable is used to predict then it is called simple linear regression and if two or more independent variables are used to predict then it is called Multiple linear regression.

Linear Regression representation

$\mathbf{y} = \mathbf{X}\cdot\mathbf{b}+\mathbf{e}$

$\mathbf{X}= \mathbf{1}, \mathbf{x_1}, \mathbf{x_2}, \mathbf{x_p}$, $\mathbf{b}= \{b_0, b_1, b_2, bp\}^T$

$\mathbf{y}= \{y_1, y_2, y_n\}^T$, $\mathbf{e}= \{e_1, e_2, e_n\}^T$

The Multiple Linear Regression representation

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + . . . + \beta_k x_k + e$, $e \sim N(0, \sigma)$ where

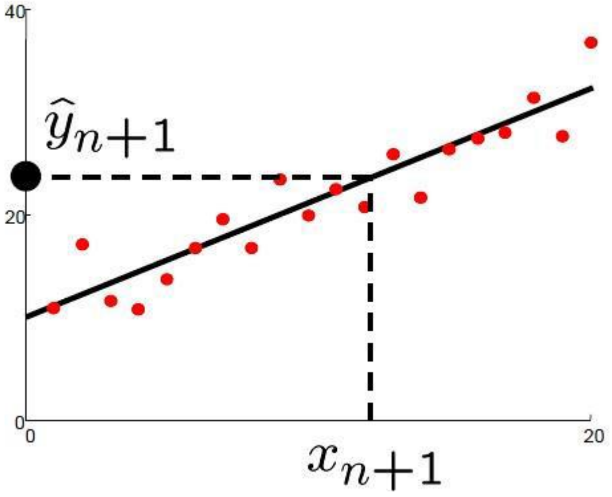
- y is the response (dependent) variable
- k is the number of predictors ˆ
- x1, x2, . . ., xk are the predictor (independent) variables
- The mean of the response, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + . . . + \beta_k x_k$, is the deterministic part of the model
- β_i describes the contribution of the predictor variable x_i
- e is the random error, which is assumed to be normally distributed with mean 0 and standard deviation σ.

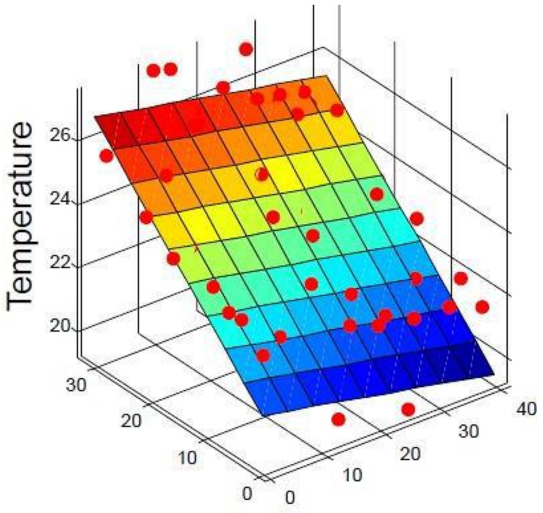
Independent variables are independent of each other. Otherwise, it causes a multicollinearity problem, two or more predictor variables are highly correlated. Should remove them.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.

There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

(b) How it does:





Samples with ONE independent variable **Samples with TWO Independent Variables**

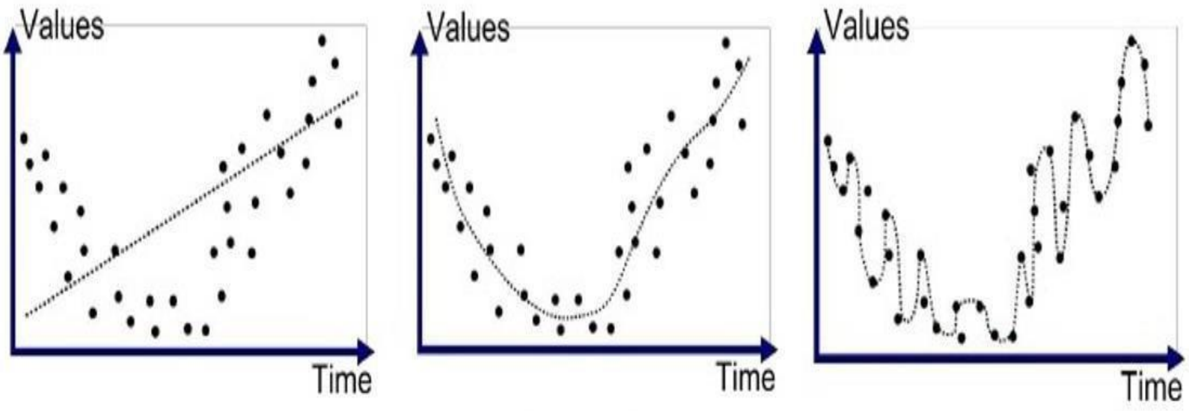
Overfitting occurs when a model is too complex and has higher-order terms

•Underfitting occurs when a model is too simple and has no Higher-order terms

•Over fit model captures the details and noise in training data

• Over fit does not capture the general trend.

•Over fit model is outstanding on training data but performs poorly on new data.



Underfitted Good Fit/Robust Overfitted

Underfitted:

$$\boxed{y = b_0 + b_1 x}$$

Good fit/Robust:

$$\boxed{y = b_0 + b_1 x + b_2 x^2 + b_3 x^3}$$

Over Fitted:

$$\boxed{\begin{array}{l} y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \cdots + b_{10} x^{10} \\ y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_{10} x_{10} \end{array}}$$

(c) Applications:

Understanding linear regression is important because it provides a scientific calculation for identifying and predicting future outcomes.

Some applications of linear regression are

i) The ability to find predictions and evaluate them can help provide benefits to many businesses and individuals, like optimized operations and detailed research materials.

ii) forecasting future opportunities and threats.

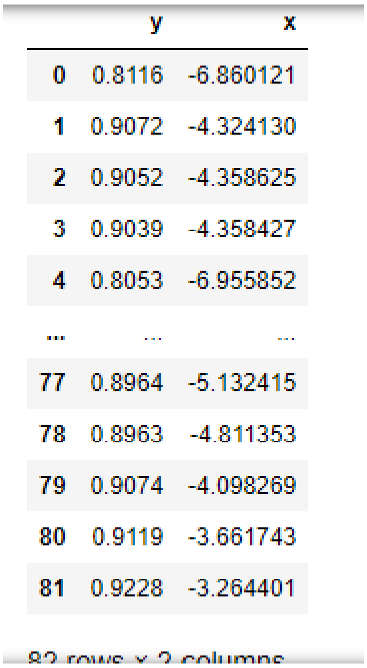
iii) Regression is a statistical method used in finance, investing, and other disciplines that attempt to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

iv) Predicting crop yields based on the amount of rainfall- Yield is a dependent variable while the measure of precipitation is an independent variable.

v) Predicting the Salary of a person based on years of experience- Therefore, Experience becomes the independent while Salary turns into the dependent variable.

RESULTS for Linear regression with higher-order terms

a)

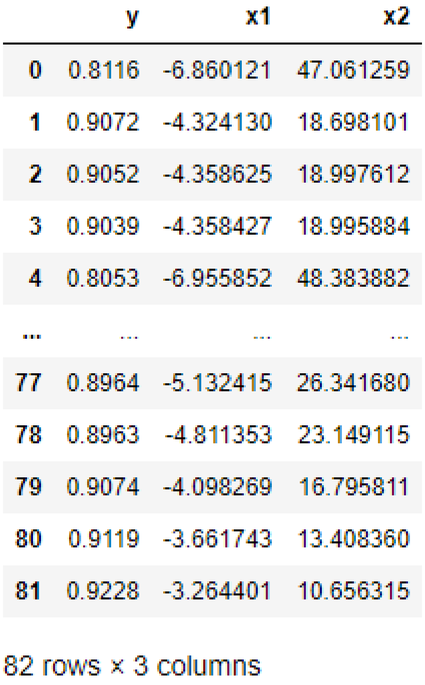


	y	x
0	0.8116	-6.860121
1	0.9072	-4.324130
2	0.9052	-4.358625
3	0.9039	-4.358427
4	0.8053	-6.955852
...	...	...
77	0.8964	-5.132415
78	0.8963	-4.811353
79	0.9074	-4.098269
80	0.9119	-3.661743
81	0.9228	-3.264401

82 rows × 2 columns

b) Data set for N = 1

a)

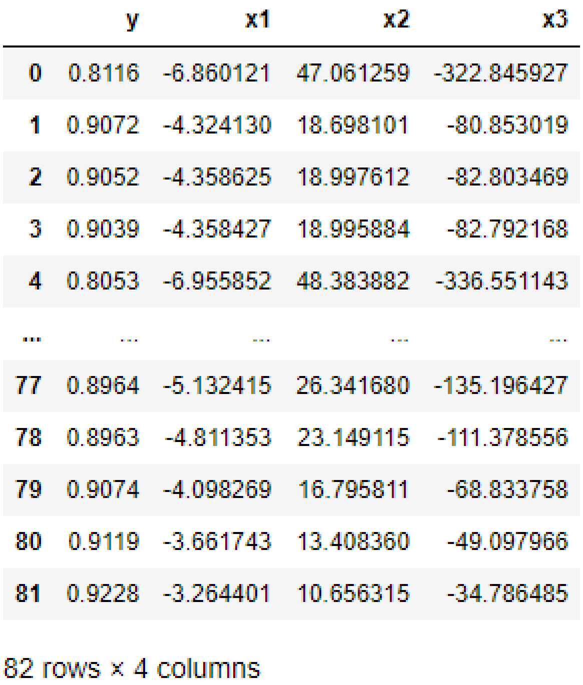


	y	x1	x2
0	0.8116	-6.860121	47.061259
1	0.9072	-4.324130	18.698101
2	0.9052	-4.358625	18.997612
3	0.9039	-4.358427	18.995884
4	0.8053	-6.955852	48.383882
...	...	...	...
77	0.8964	-5.132415	26.341680
78	0.8963	-4.811353	23.149115
79	0.9074	-4.098269	16.795811
80	0.9119	-3.661743	13.408360
81	0.9228	-3.264401	10.656315

82 rows × 3 columns

b) Data set for N = 2

a)

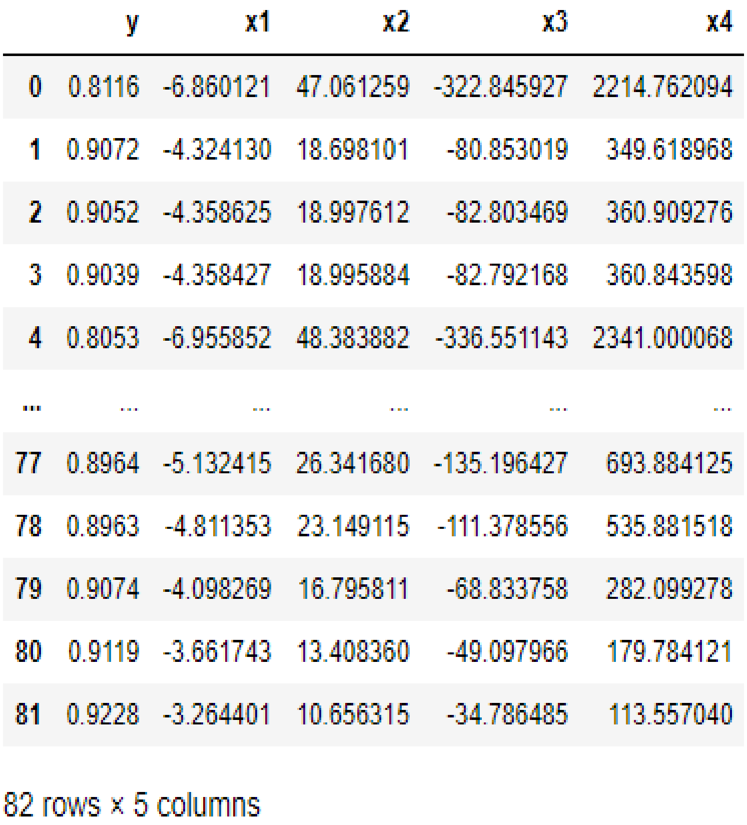


	y	x1	x2	x3
0	0.8116	-6.860121	47.061259	-322.845927
1	0.9072	-4.324130	18.698101	-80.853019
2	0.9052	-4.358625	18.997612	-82.803469
3	0.9039	-4.358427	18.995884	-82.792168
4	0.8053	-6.955852	48.383882	-336.551143
...	...	...	...	...
77	0.8964	-5.132415	26.341680	-135.196427
78	0.8963	-4.811353	23.149115	-111.378556
79	0.9074	-4.098269	16.795811	-68.833758
80	0.9119	-3.661743	13.408360	-49.097966
81	0.9228	-3.264401	10.656315	-34.786485

82 rows × 4 columns

b) Data set for N = 3

a)

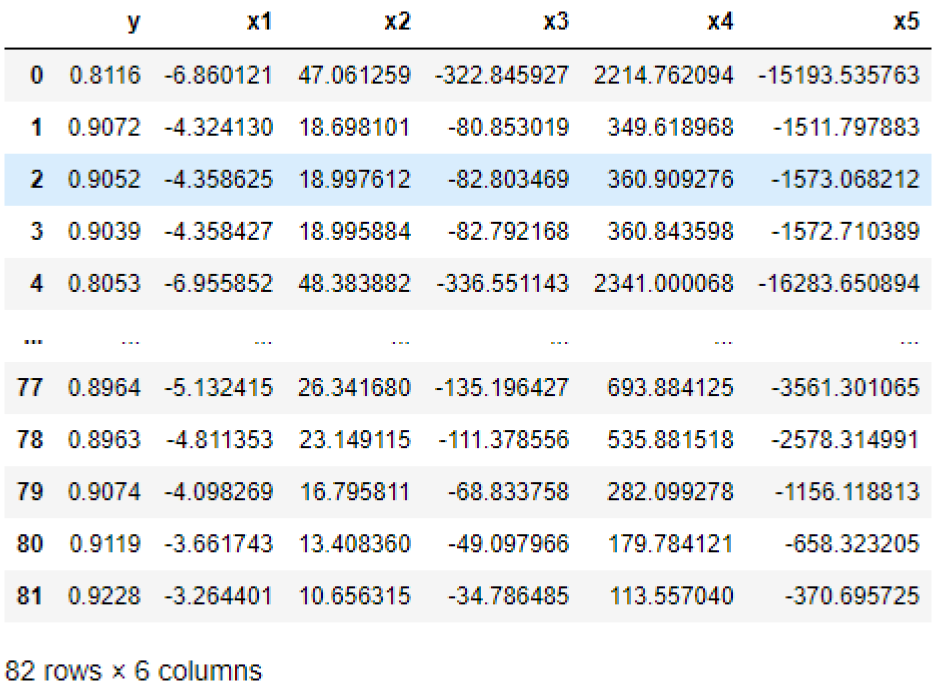


	y	x1	x2	x3	x4
0	0.8116	-6.860121	47.061259	-322.845927	2214.762094
1	0.9072	-4.324130	18.698101	-80.853019	349.618968
2	0.9052	-4.358625	18.997612	-82.803469	360.909276
3	0.9039	-4.358427	18.995884	-82.792168	360.843598
4	0.8053	-6.955852	48.383882	-336.551143	2341.000068
...	...	...	...	...	...
77	0.8964	-5.132415	26.341680	-135.196427	693.884125
78	0.8963	-4.811353	23.149115	-111.378556	535.881518
79	0.9074	-4.098269	16.795811	-68.833758	282.099278
80	0.9119	-3.661743	13.408360	-49.097966	179.784121
81	0.9228	-3.264401	10.656315	-34.786485	113.557040

82 rows × 5 columns

b) Data set for N = 4

a)

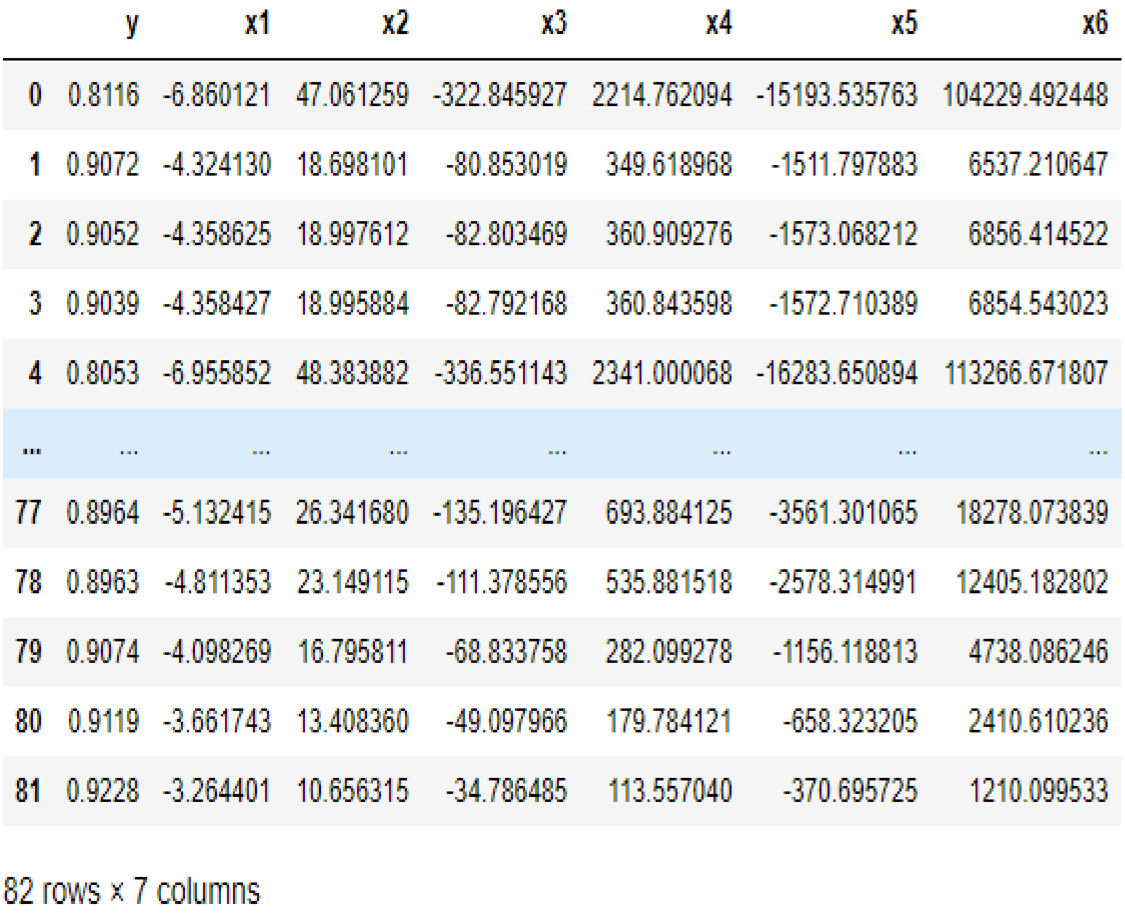


	y	x1	x2	x3	x4	x5
0	0.8116	-6.860121	47.061259	-322.845927	2214.762094	-15193.535763
1	0.9072	-4.324130	18.698101	-80.853019	349.618968	-1511.797883
2	0.9052	-4.358625	18.997612	-82.803469	360.909276	-1573.068212
3	0.9039	-4.358427	18.995884	-82.792168	360.843598	-1572.710389
4	0.8053	-6.955852	48.383882	-336.551143	2341.000068	-16283.650894
...	...	...	...	...	...	...
77	0.8964	-5.132415	26.341680	-135.196427	693.884125	-3561.301065
78	0.8963	-4.811353	23.149115	-111.378556	535.881518	-2578.314991
79	0.9074	-4.098269	16.795811	-68.833758	282.099278	-1156.118813
80	0.9119	-3.661743	13.408360	-49.097966	179.784121	-658.323205
81	0.9228	-3.264401	10.656315	-34.786485	113.557040	-370.695725

82 rows × 6 columns

b) Data set for N = 5

a)

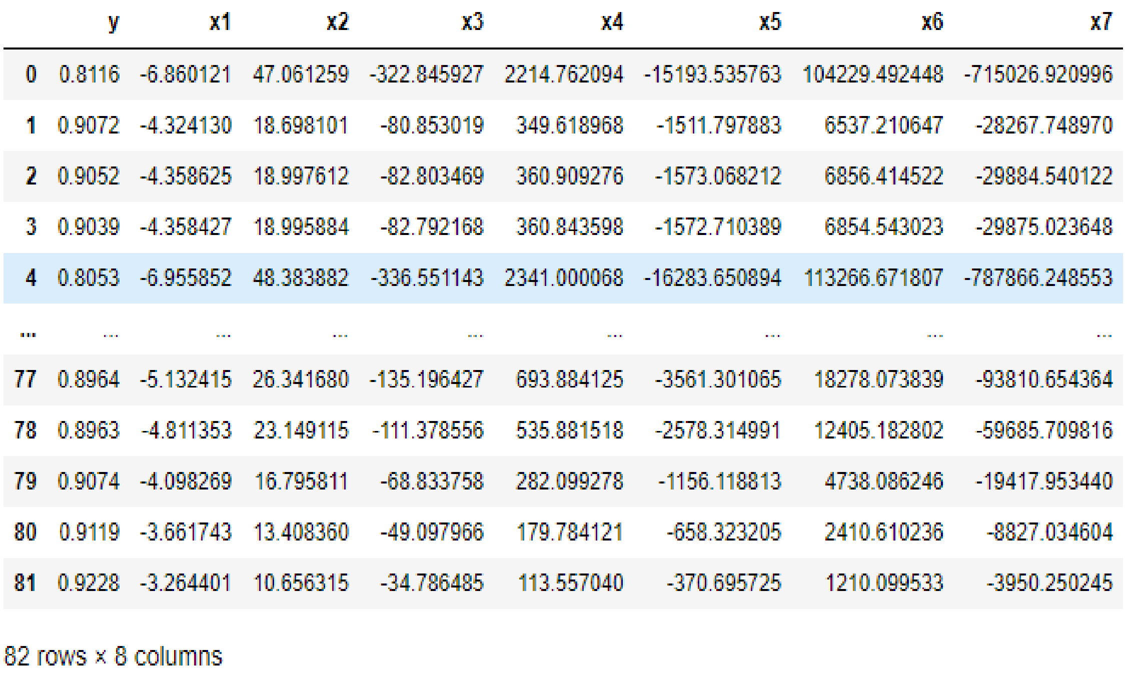


	y	x1	x2	x3	x4	x5	x6
0	0.8116	-6.860121	47.061259	-322.845927	2214.762094	-15193.535763	104229.492448
1	0.9072	-4.324130	18.698101	-80.853019	349.618968	-1511.797883	6537.210647
2	0.9052	-4.358625	18.997612	-82.803469	360.909276	-1573.068212	6856.414522
3	0.9039	-4.358427	18.995884	-82.792168	360.843598	-1572.710389	6854.543023
4	0.8053	-6.955852	48.383882	-336.551143	2341.000068	-16283.650894	113266.671807
...	...	...	...	...	...	...	...
77	0.8964	-5.132415	26.341680	-135.196427	693.884125	-3561.301065	18278.073839
78	0.8963	-4.811353	23.149115	-111.378556	535.881518	-2578.314991	12405.182802
79	0.9074	-4.098269	16.795811	-68.833758	282.099278	-1156.118813	4738.086246
80	0.9119	-3.661743	13.408360	-49.097966	179.784121	-658.323205	2410.610236
81	0.9228	-3.264401	10.656315	-34.786485	113.557040	-370.695725	1210.099533

82 rows × 7 columns

b) Data set for N = 6

a)

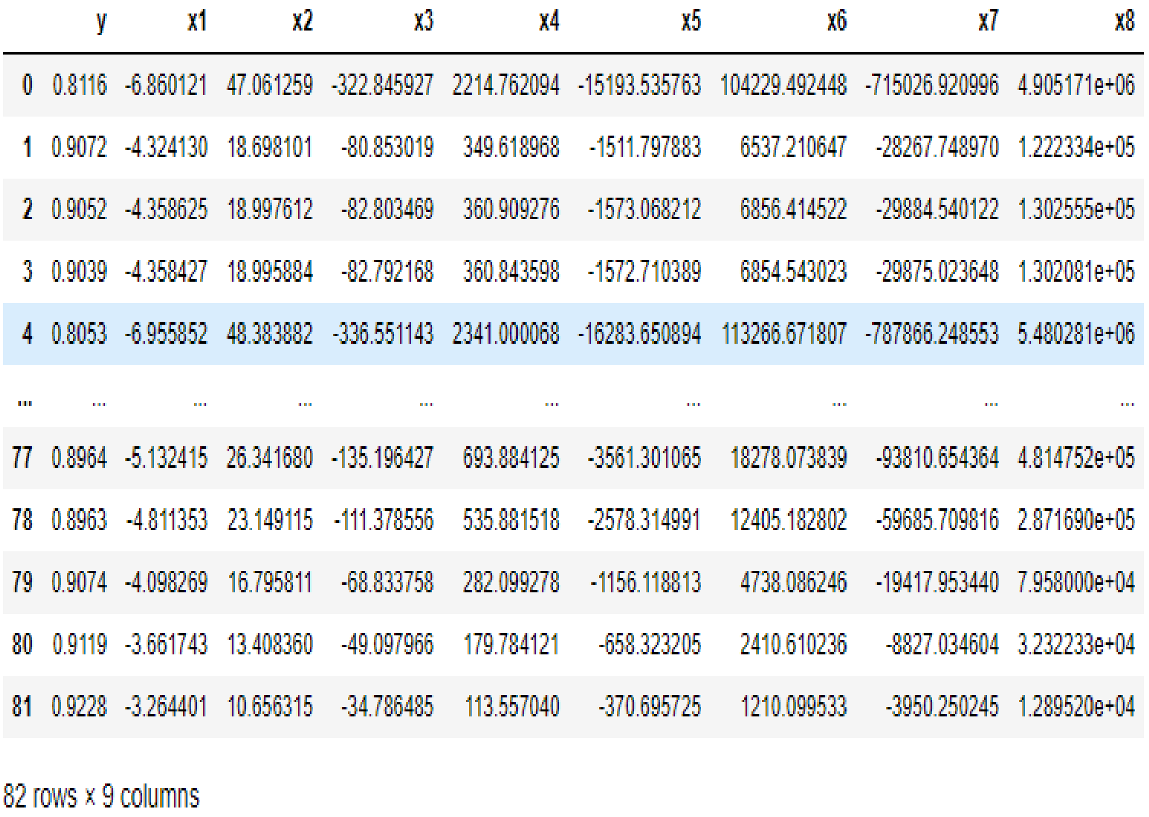


	y	x1	x2	x3	x4	x5	x6	x7
0	0.8116	-6.860121	47.061259	-322.845927	2214.762094	-15193.535763	104229.492448	-715026.920996
1	0.9072	-4.324130	18.698101	-80.853019	349.618968	-1511.797883	6537.210647	-28267.748970
2	0.9052	-4.358625	18.997612	-82.803469	360.909276	-1573.068212	6856.414522	-29884.540122
3	0.9039	-4.358427	18.995884	-82.792168	360.843598	-1572.710389	6854.543023	-29875.023648
4	0.8053	-6.955852	48.383882	-336.551143	2341.000068	-16283.650894	113266.671807	-787866.248553
...	...	...	...	...	...	...	...	...
77	0.8964	-5.132415	26.341680	-135.196427	693.884125	-3561.301065	18278.073839	-93810.654364
78	0.8963	-4.811353	23.149115	-111.378556	535.881518	-2578.314991	12405.182802	-59685.709816
79	0.9074	-4.098269	16.795811	-68.833758	282.099278	-1156.118813	4738.086246	-19417.953440
80	0.9119	-3.661743	13.408360	-49.097966	179.784121	-658.323205	2410.610236	-8827.034604
81	0.9228	-3.264401	10.656315	-34.786485	113.557040	-370.695725	1210.099533	-3950.250245

82 rows × 8 columns

b) Data set for N = 7

a)

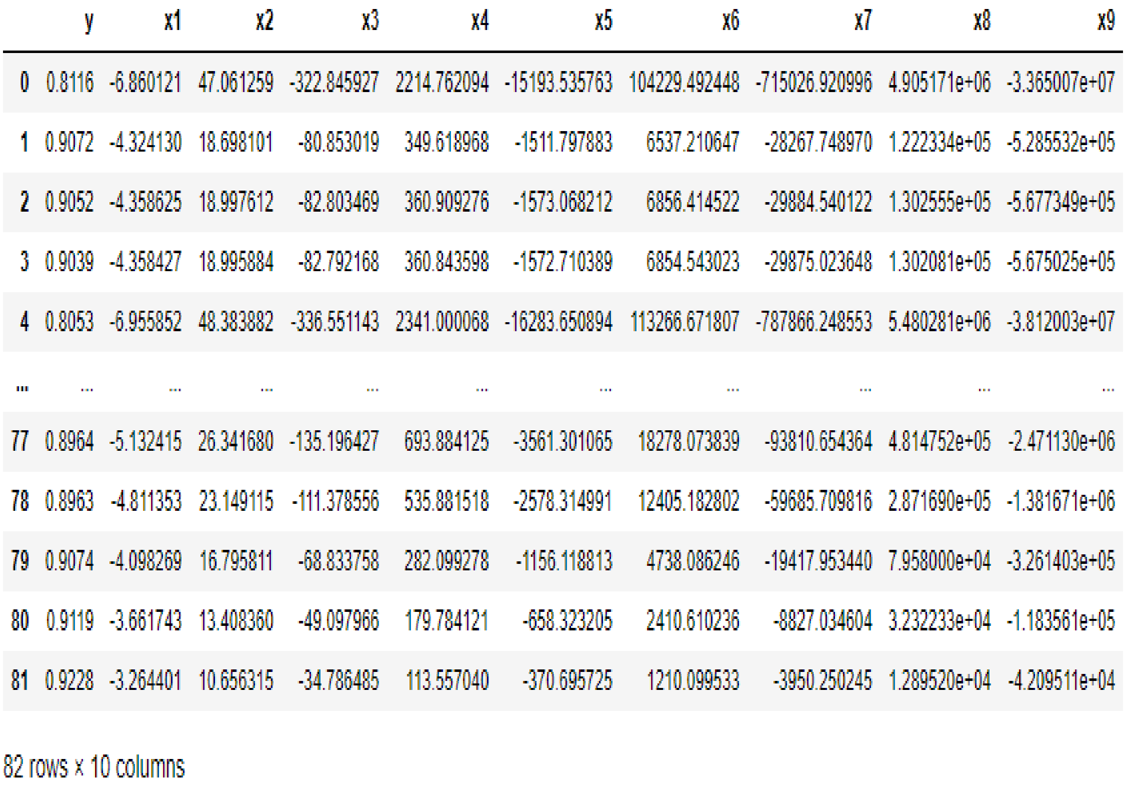


	y	x1	x2	x3	x4	x5	x6	x7	x8
0	0.8116	-6.860121	47.061259	-322.845927	2214.762094	-15193.535763	104229.492448	-715026.920996	4.905171e+06
1	0.9072	-4.324130	18.698101	-80.853019	349.618968	-1511.797883	6537.210647	-28267.748970	1.222334e+05
2	0.9052	-4.358625	18.997612	-82.803469	360.909276	-1573.068212	6856.414522	-29884.540122	1.302555e+05
3	0.9039	-4.358427	18.995884	-82.792168	360.843598	-1572.710389	6854.543023	-29875.023648	1.302081e+05
4	0.8053	-6.955852	48.383882	-336.551143	2341.000068	-16283.650894	113266.671807	-787866.248553	5.480281e+06
...	...	...	...	...	...	...	...	...	...
77	0.8964	-5.132415	26.341680	-135.196427	693.884125	-3561.301065	18278.073839	-93810.654364	4.814752e+05
78	0.8963	-4.811353	23.149115	-111.378556	535.881518	-2578.314991	12405.182802	-59685.709816	2.871690e+05
79	0.9074	-4.098269	16.795811	-68.833758	282.099278	-1156.118813	4738.086246	-19417.953440	7.958000e+04
80	0.9119	-3.661743	13.408360	-49.097966	179.784121	-658.323205	2410.610236	-8827.034604	3.232233e+04
81	0.9228	-3.264401	10.656315	-34.786485	113.557040	-370.695725	1210.099533	-3950.250245	1.289520e+04

82 rows × 9 columns

b) Data set for N = 8

a)

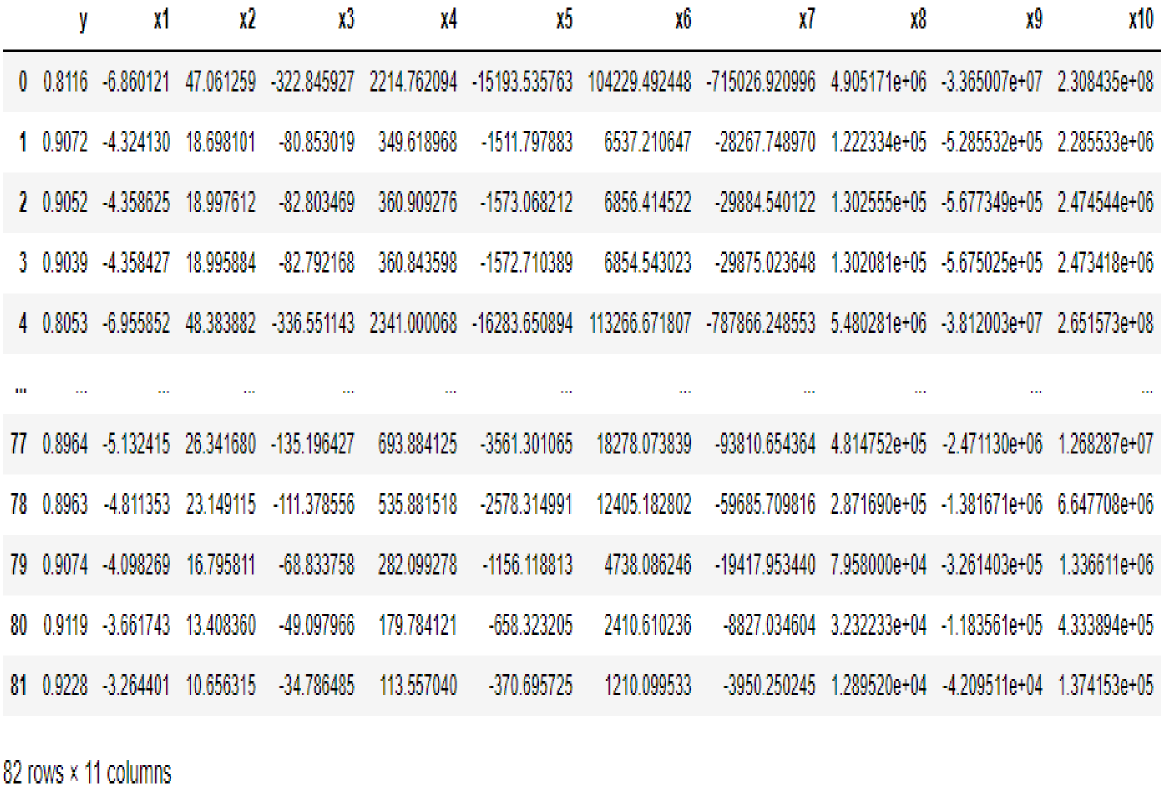


	y	x1	x2	x3	x4	x5	x6	x7	x8	x9
0	0.8116	-6.860121	47.061259	-322.845927	2214.762094	-15193.535763	104229.492448	-715026.920996	4.905171e+06	-3.365007e+07
1	0.9072	-4.324130	18.698101	-80.853019	349.618968	-1511.797883	6537.210647	-28267.748970	1.222334e+05	-5.285532e+05
2	0.9052	-4.358625	18.997612	-82.803469	360.909276	-1573.068212	6856.414522	-29884.540122	1.302555e+05	-5.677349e+05
3	0.9039	-4.358427	18.995884	-82.792168	360.843598	-1572.710389	6854.543023	-29875.023648	1.302081e+05	-5.675025e+05
4	0.8053	-6.955852	48.383882	-336.551143	2341.000068	-16283.650894	113266.671807	-787866.248553	5.480281e+06	-3.812003e+07
...	...	...	...	...	...	...	...	...	...	...
77	0.8964	-5.132415	26.341680	-135.196427	693.884125	-3561.301065	18278.073839	-93810.654364	4.814752e+05	-2.471130e+06
78	0.8963	-4.811353	23.149115	-111.378556	535.881518	-2578.314991	12405.182802	-59685.709816	2.871690e+05	-1.381671e+06
79	0.9074	-4.098269	16.795811	-68.833758	282.099278	-1156.118813	4738.086246	-19417.953440	7.958000e+04	-3.261403e+05
80	0.9119	-3.661743	13.408360	-49.097966	179.784121	-658.323205	2410.610236	-8827.034604	3.232233e+04	-1.183561e+05
81	0.9228	-3.264401	10.656315	-34.786485	113.557040	-370.695725	1210.099533	-3950.250245	1.289520e+04	-4.209511e+04

82 rows × 10 columns

b) Data set for N=9

a)

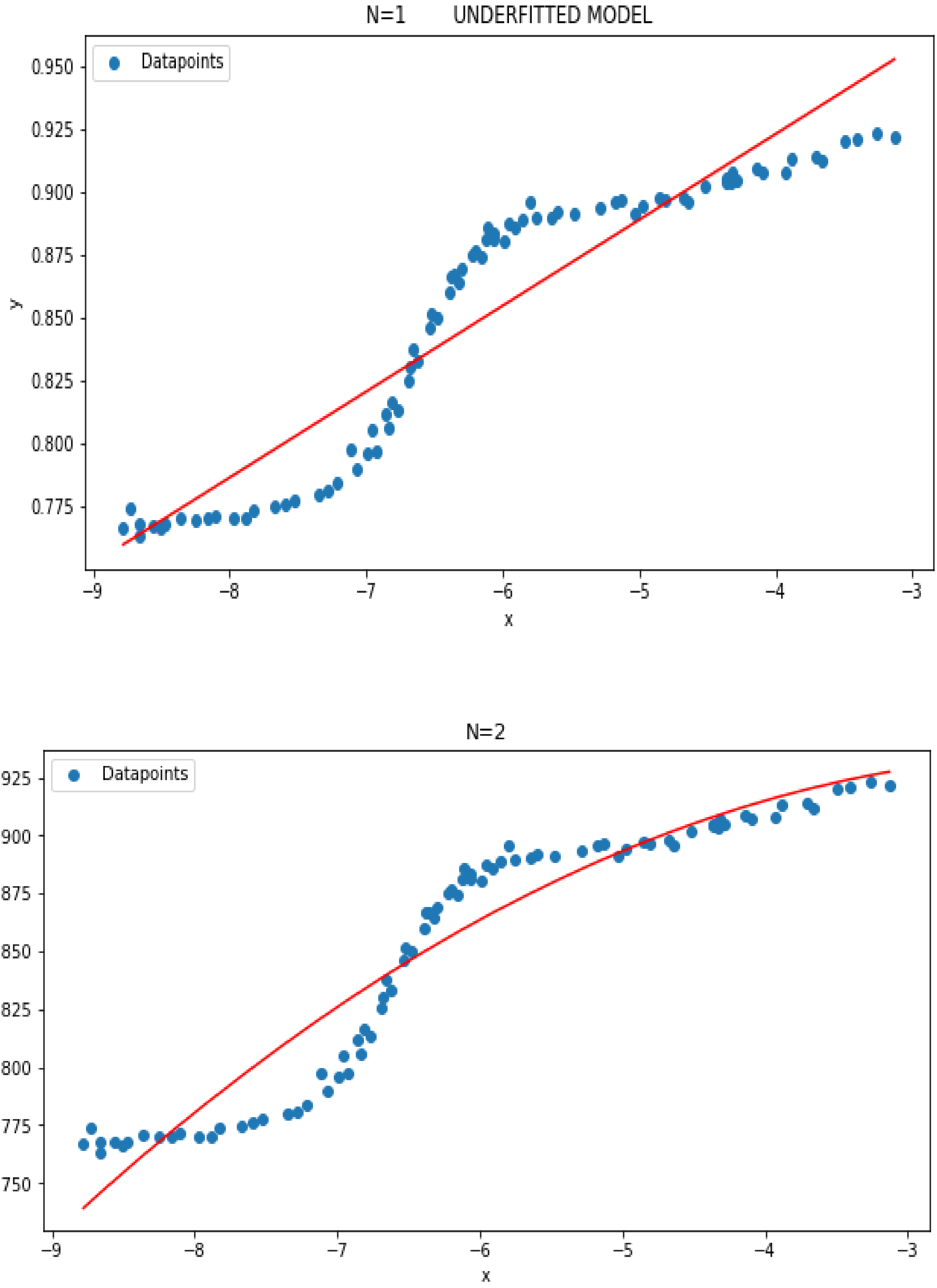


	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
0	0.8116	-6.860121	47.061259	-322.845927	2214.762094	-15193.535763	104229.492448	-715026.920996	4.905171e+06	-3.365007e+07	2.308435e+08
1	0.9072	-4.324130	18.698101	-80.853019	349.618968	-1511.797883	6537.210647	-28267.748970	1.222334e+05	-5.285532e+05	2.285533e+06
2	0.9052	-4.358625	18.997612	-82.803469	360.909276	-1573.068212	6856.414522	-29884.540122	1.302555e+05	-5.677349e+05	2.474544e+06
3	0.9039	-4.358427	18.995884	-82.792168	360.843598	-1572.710389	6854.543023	-29875.023648	1.302081e+05	-5.675025e+05	2.473418e+06
4	0.8053	-6.955852	48.383882	-336.551143	2341.000068	-16283.650894	113266.671807	-787866.248553	5.480281e+06	-3.812003e+07	2.651573e+08
...	...	...	...	...	...	...	...	...	...	...	...
77	0.8964	-5.132415	26.341680	-135.196427	693.884125	-3561.301065	18278.073839	-93810.654364	4.814752e+05	-2.471130e+06	1.268287e+07
78	0.8963	-4.811353	23.149115	-111.378556	535.881518	-2578.314991	12405.182802	-59685.709816	2.871690e+05	-1.381671e+06	6.647708e+06
79	0.9074	-4.098269	16.795811	-68.833758	282.099278	-1156.118813	4738.086246	-19417.953440	7.958000e+04	-3.261403e+05	1.336611e+06
80	0.9119	-3.661743	13.408360	-49.097966	179.784121	-658.323205	2410.610236	-8827.034604	3.232233e+04	-1.183561e+05	4.333894e+05
81	0.9228	-3.264401	10.656315	-34.786485	113.557040	-370.695725	1210.099533	-3950.250245	1.289520e+04	-4.209511e+04	1.374153e+05

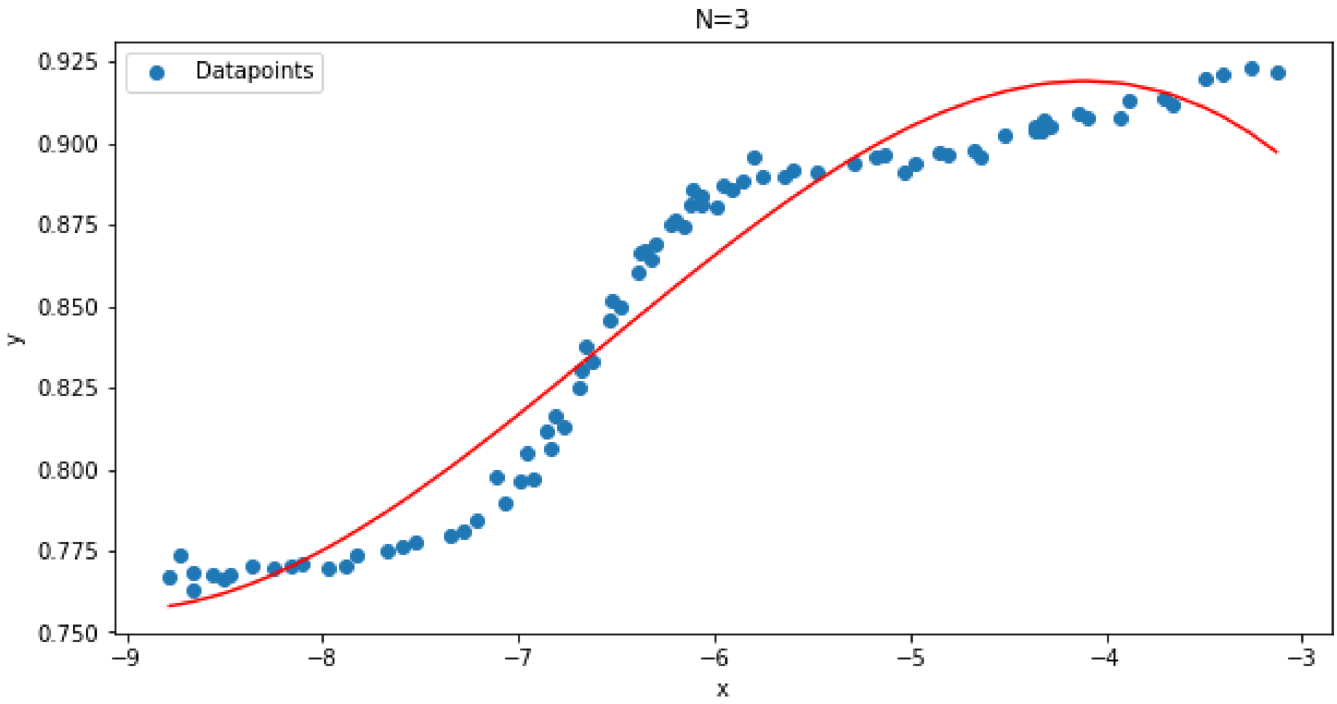
82 rows × 11 columns

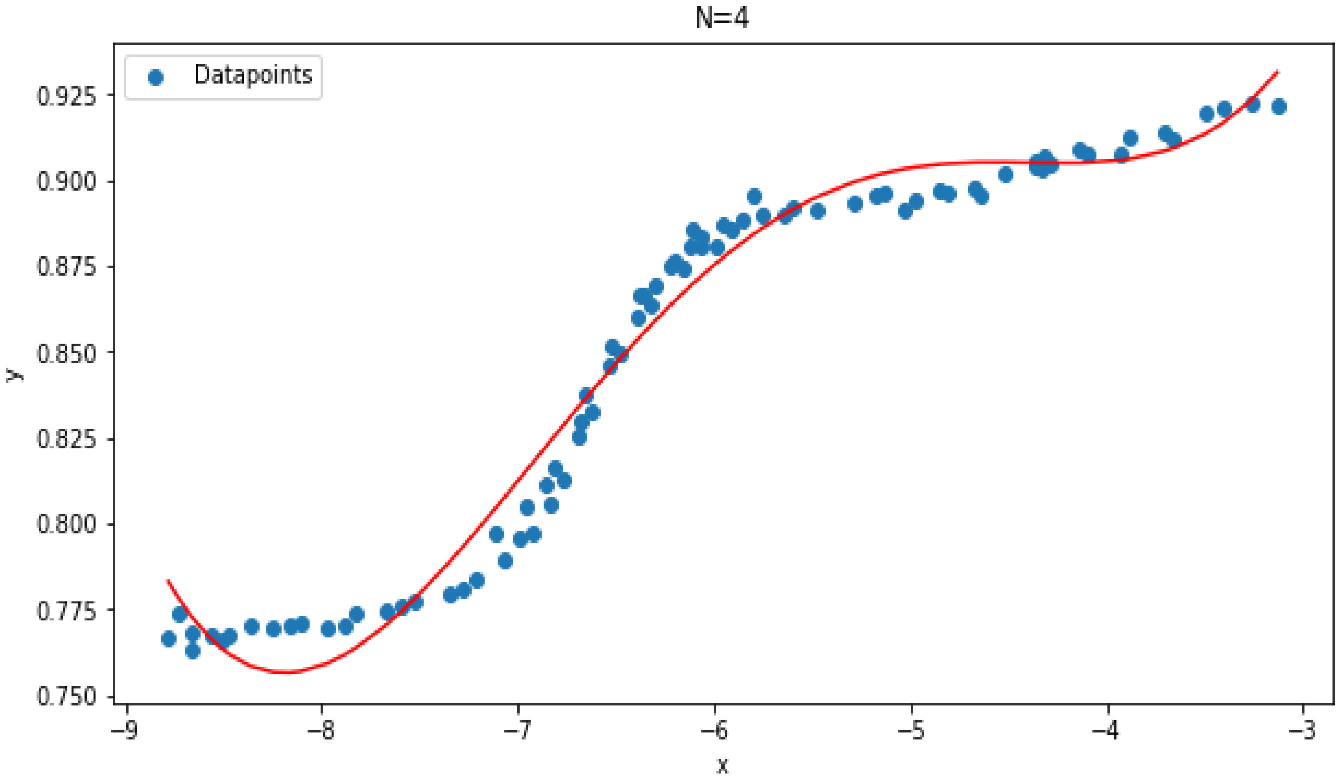
b) Data set for N=10

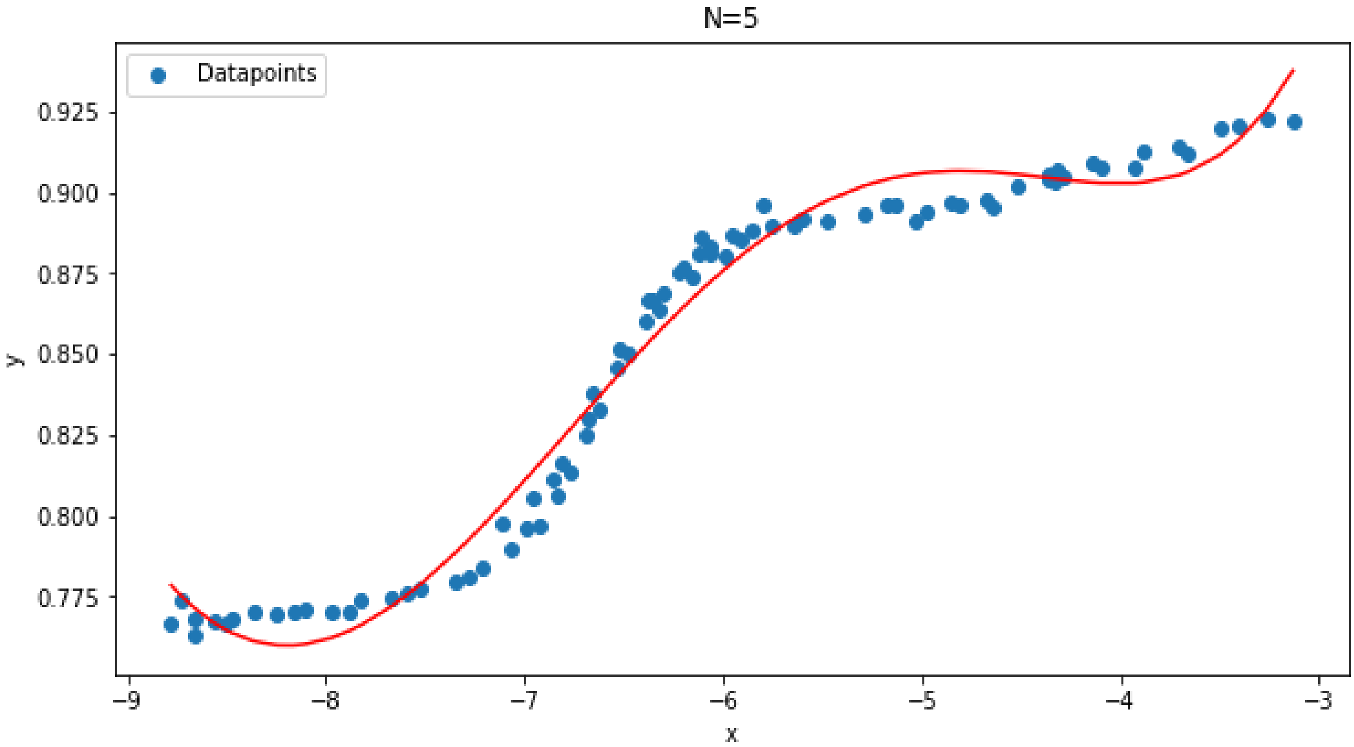
a)

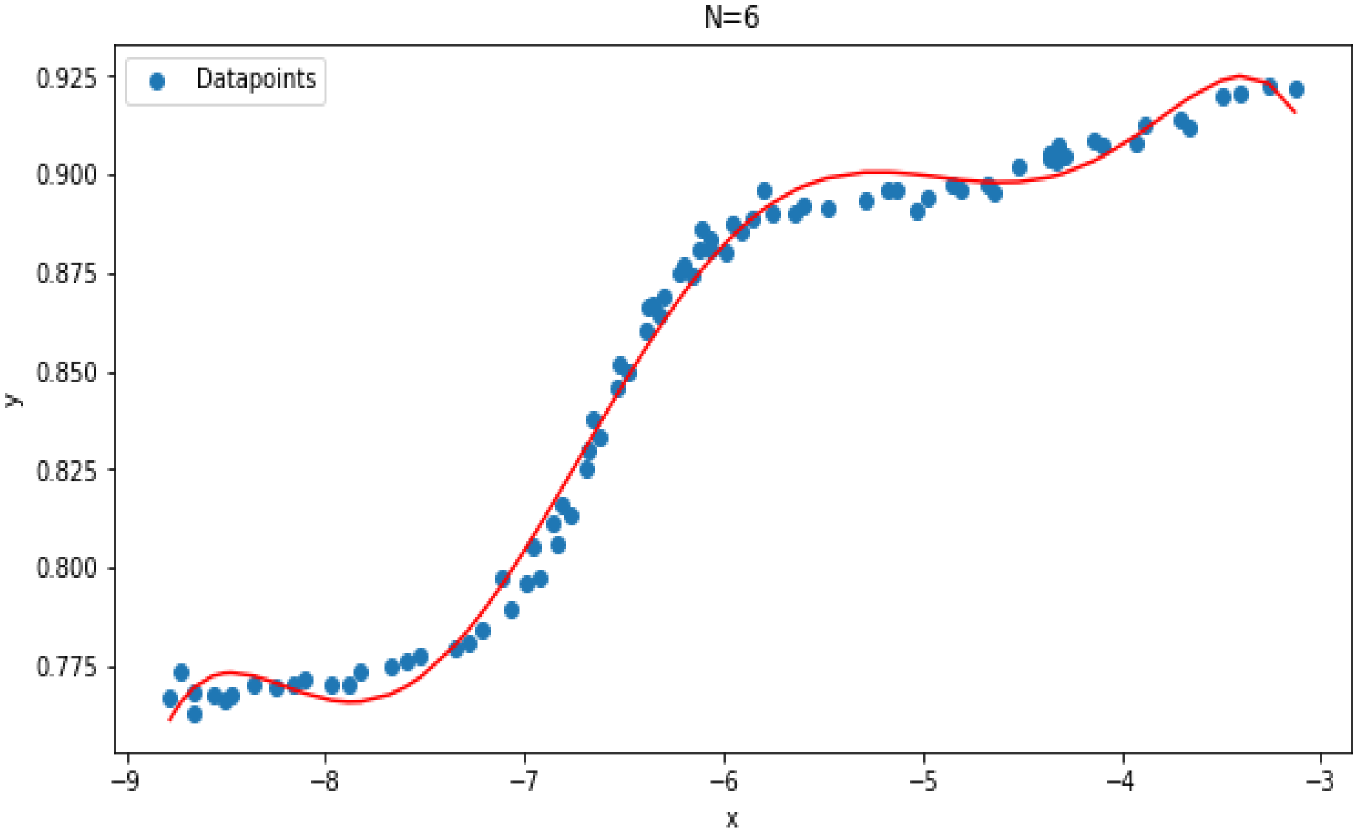

N=1 UNDERFITTED MODEL

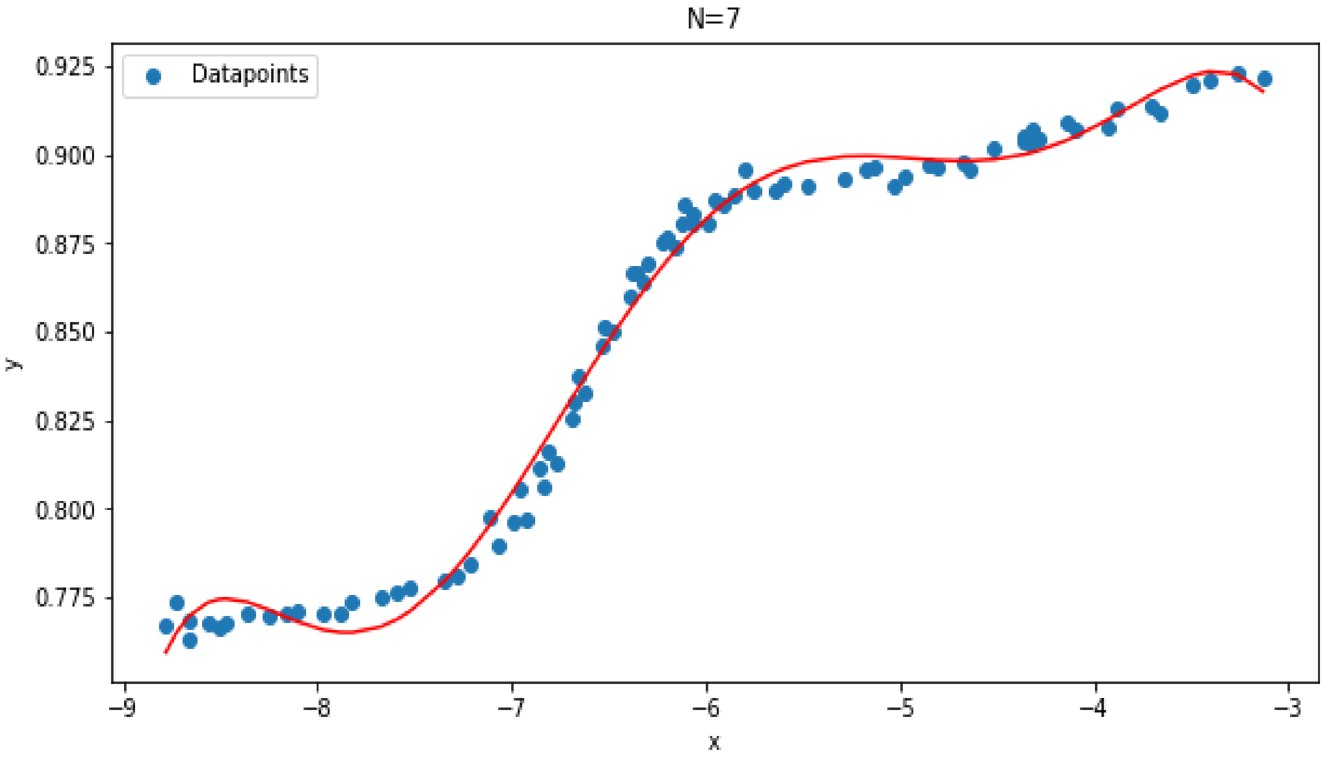
N=2

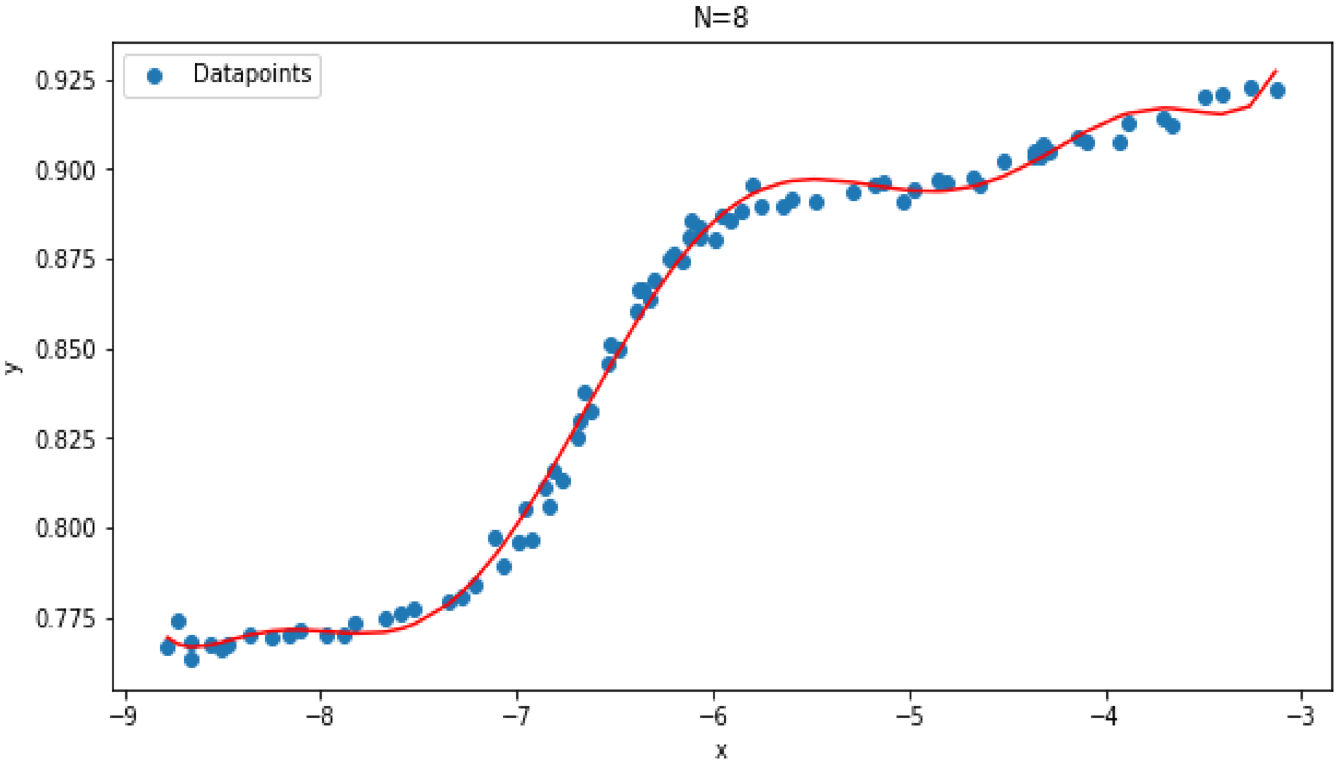

N=3
Datapoints
0.925
0.900
0.875
0.850
0.825
0.800
0.775
0.750
y
−9
−8
−7
−6
−5
−4
−3
x

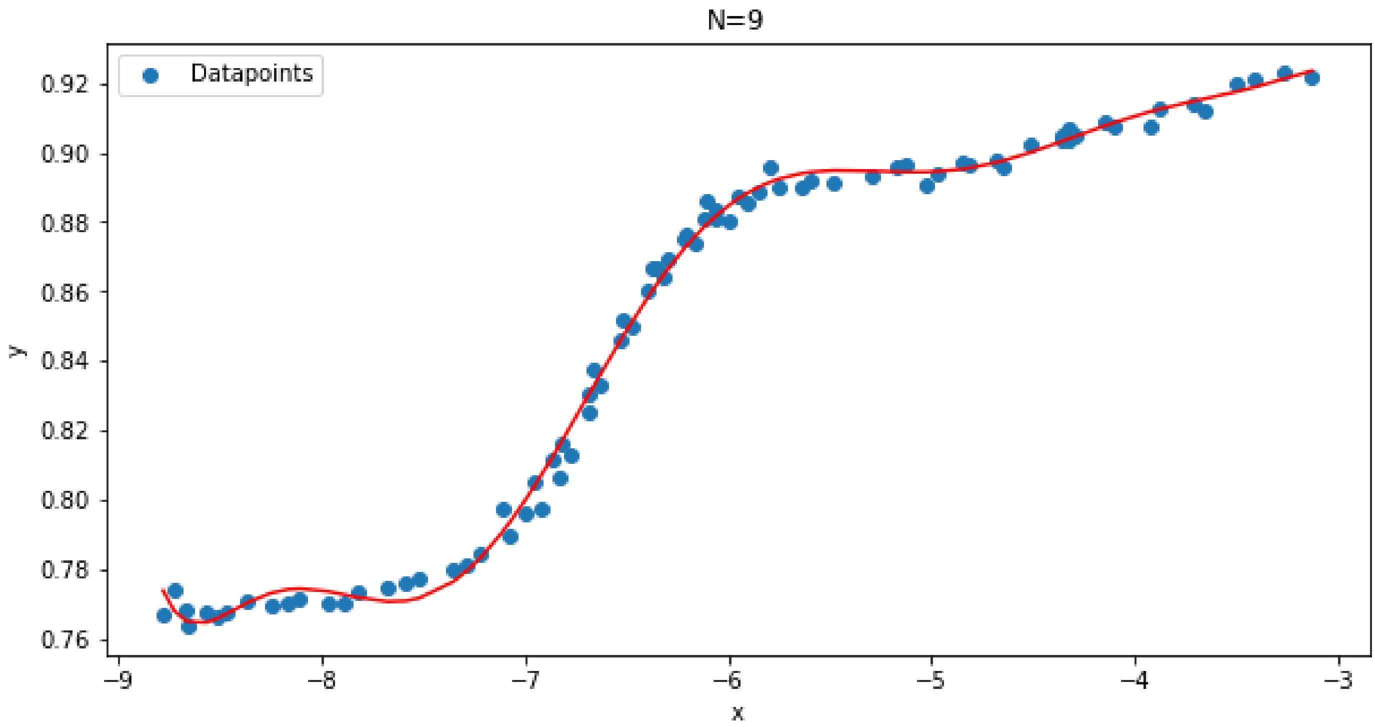

N=4
Datapoints
0.925
0.900
0.875
0.850
0.825
0.800
0.775
0.750
y
−9
−8
−7
−6
−5
−4
−3
x

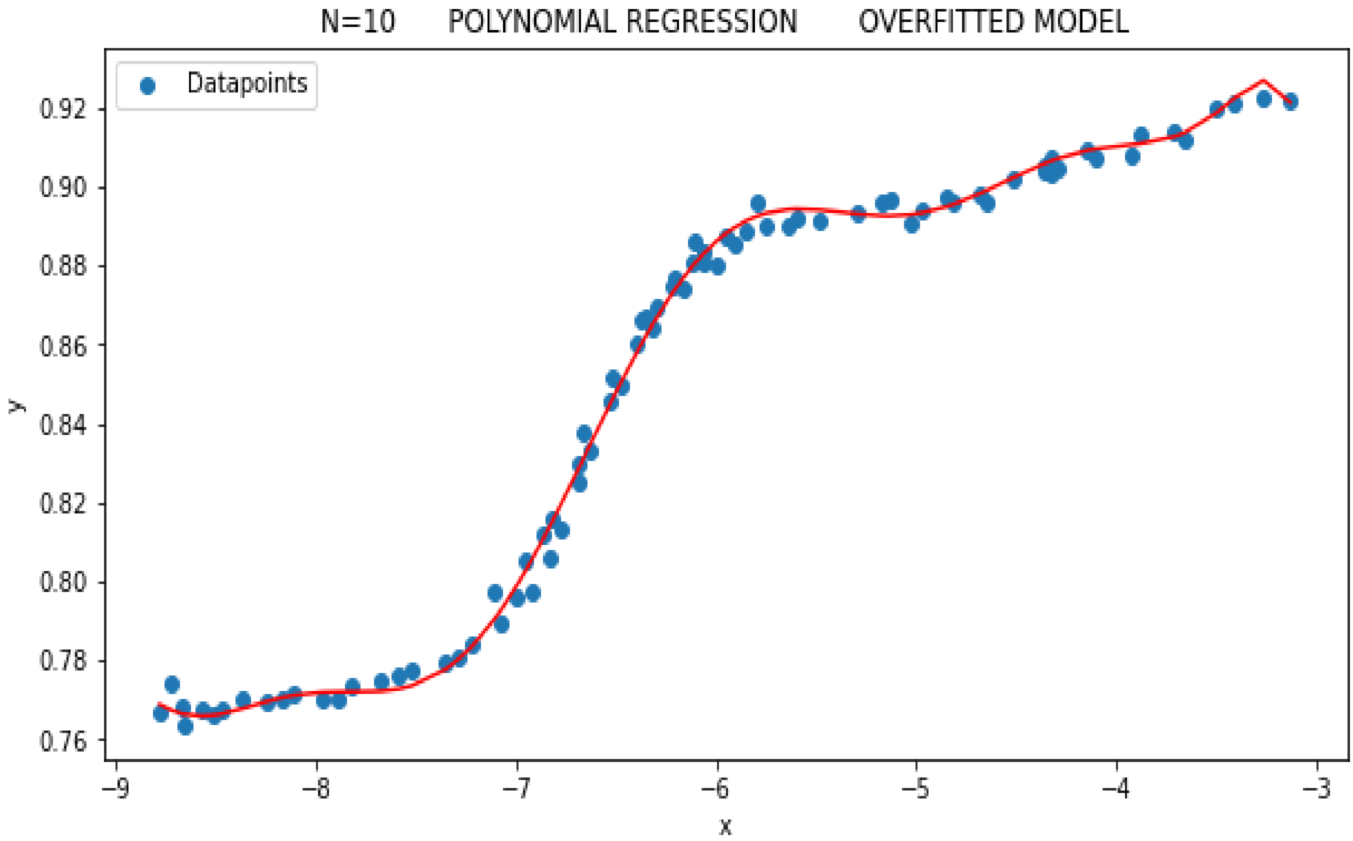

N=5
Datapoints
0.925
0.900
0.875
0.850
0.825
0.800
0.775
y
−9
−8
−7
−6
−5
−4
−3
x


N=6
Datapoints
0.925
0.900
0.875
0.850
0.825
0.800
0.775
y
−9
−8
−7
−6
−5
−4
−3
x


N=7
Datapoints
0.925
0.900
0.875
0.850
0.825
0.800
0.775
y
−9
−8
−7
−6
−5
−4
−3
x


N=8
Datapoints
0.925
0.900
0.875
0.850
0.825
0.800
0.775
y
−9
−8
−7
−6
−5
−4
−3
x


N=9
Datapoints
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0.78
0.76
y
−9
−8
−7
−6
−5
−4
−3
x


N=10 POLYNOMIAL REGRESSION OVERFITTED MODEL
Datapoints
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0.78
0.76
y
−9
−8
−7
−6
−5
−4
−3
x

c) Observation: "Linear regression models assume that the residuals are normally distributed."

This statement accurately reflects a key assumption in linear regression. Linear regression models assume that the residuals, which represent the differences between the observed and predicted values, follow a normal distribution. This assumption is crucial for valid statistical inferences, hypothesis testing, and the construction of reliable confidence intervals. It ensures that the errors in prediction conform to a symmetrical and bell-shaped distribution, supporting the robustness and reliability of the linear regression analysis.

d) Conclusion:

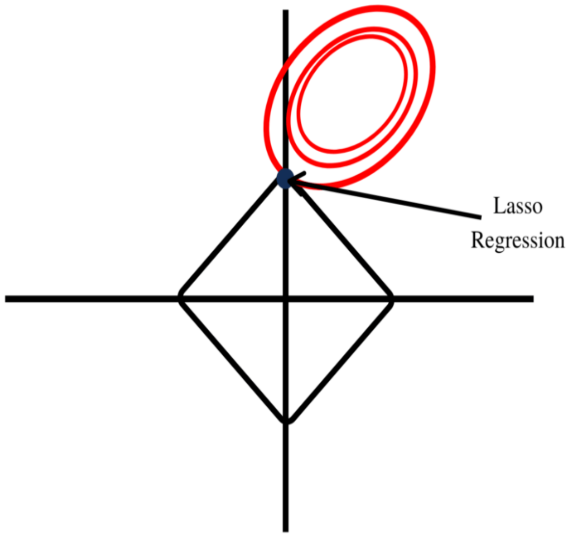
Regression analysis is indeed a supervised learning algorithm used for predicting a continuous outcome variable based on one or more predictor variables. While it can utilize labeled data, it is not restricted to such data; both labeled and unlabeled datasets are applicable for training and evaluation.

In the context of the linear regression model, there are parameters for both the slope (coefficients associated with predictor variables) and intercept. In simple linear regression, there are parameters for the slope and intercept, and in multiple linear regression, there are multiple slope parameters along with an intercept. The linear relationship assumed by the model characterizes the dependence between the dependent and independent variables.

LASSO Model

a) What it is:

LASSO, short for Least Absolute Shrinkage and Selection Operator, is a statistical formula whose main purpose is the feature selection and regularization of data models. The method was first introduced in 1996 by Statistics Professor Robert Tibshirani. LASSO introduces parameters to the sum of a model, giving it an upper bound that acts as a constraint for the sum to include absolute parameters within an allowable range.



b)How it does:

LASSO Regularization Controls the model complexity by penalizing higher-order terms in the model.

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

If a regularization term is added, the model tries to minimize both the loss and complexity of the model.

The LASSO method regularizes model parameters by shrinking the regression coefficients and reducing some of them to zero. The feature selection phase occurs after the shrinkage, where every non-zero value is selected to be used in the model. This method is significant in the minimization of prediction errors that are common in statistical models.

When to Use:

Lasso regression penalizes less important features of your dataset and makes their respective coefficients zero, thereby eliminating them. Thus it provides you with the benefit of feature selection and simple model creation.

So, if the dataset has high dimensionality and high correlation, lasso regression can be used.

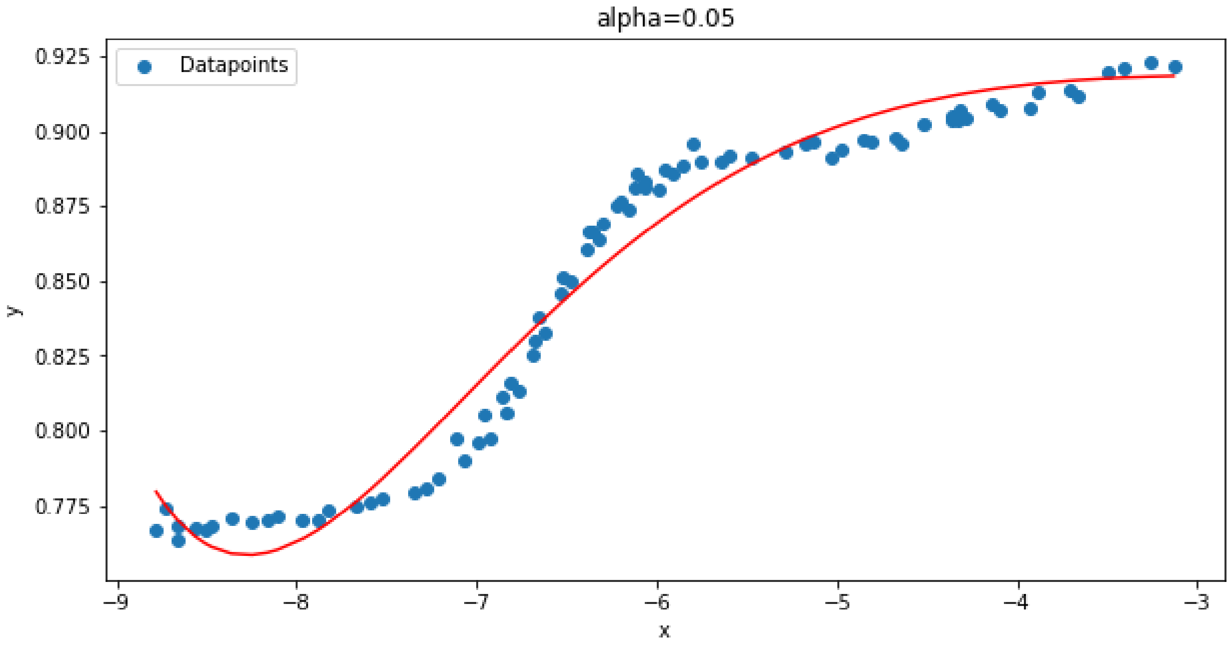
c) Applications:

Lasso regressions, in particular, are well suited for building forecasting models when the number of potential covariates is large, and the number of observations is small or roughly equal to the number of covariates.

A LASSO-based forecast model for endemic infectious diseases is proposed**.**

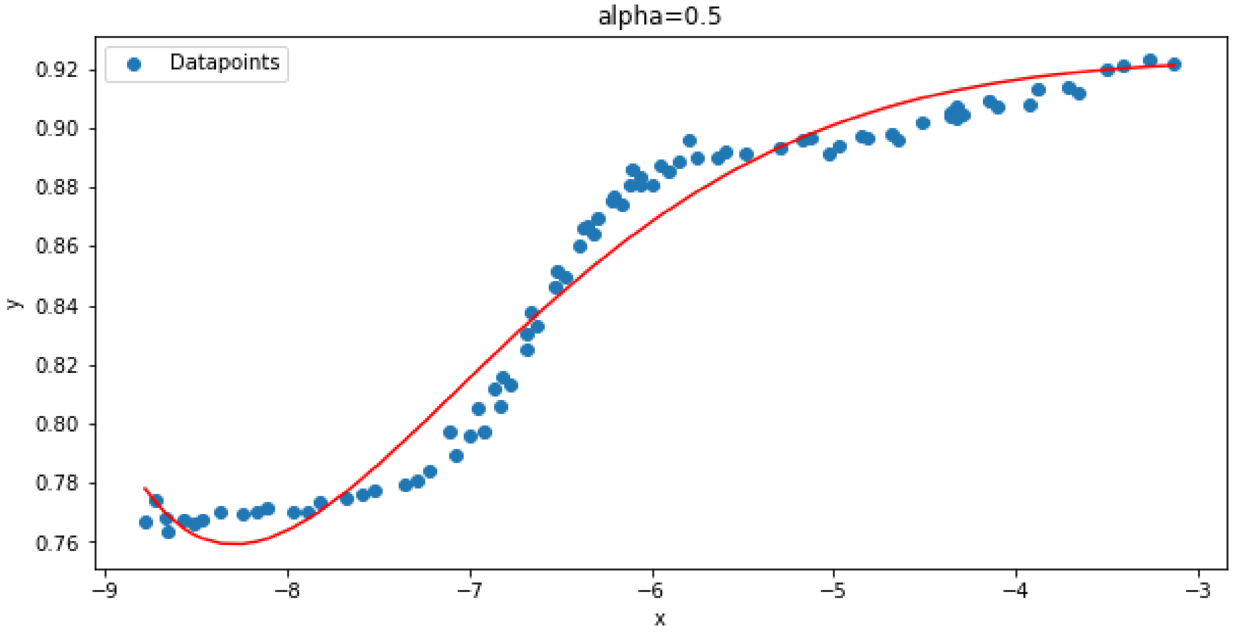
RESULTS OF LASSO Model

a)



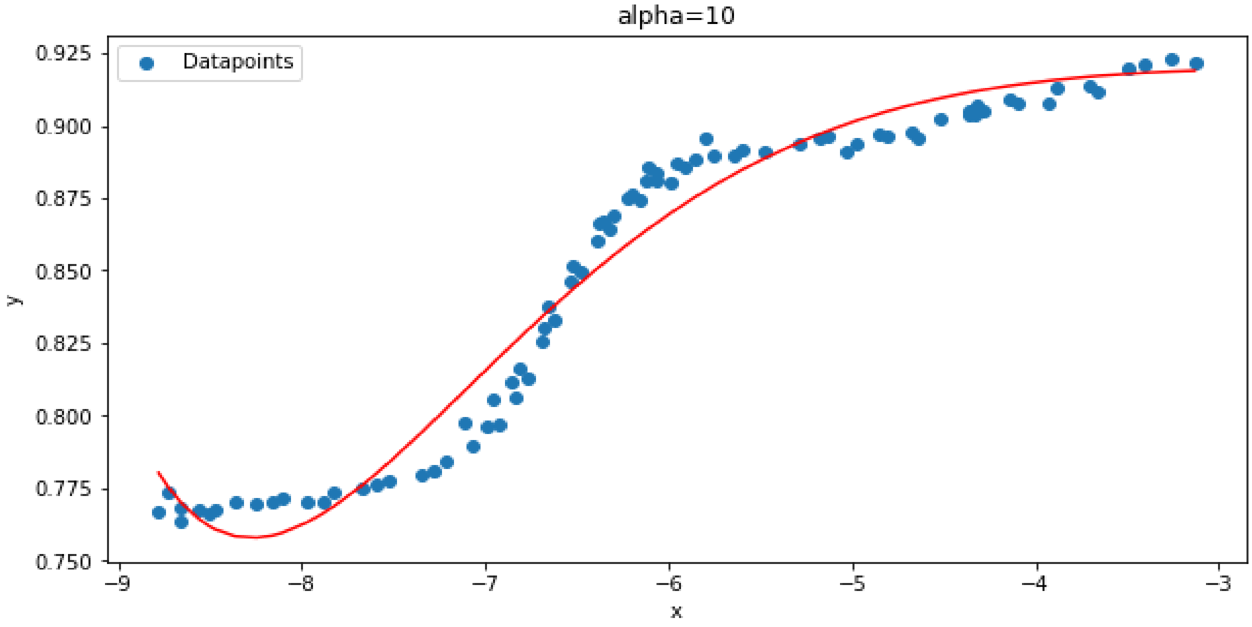
output when alpha = 0.05

b)



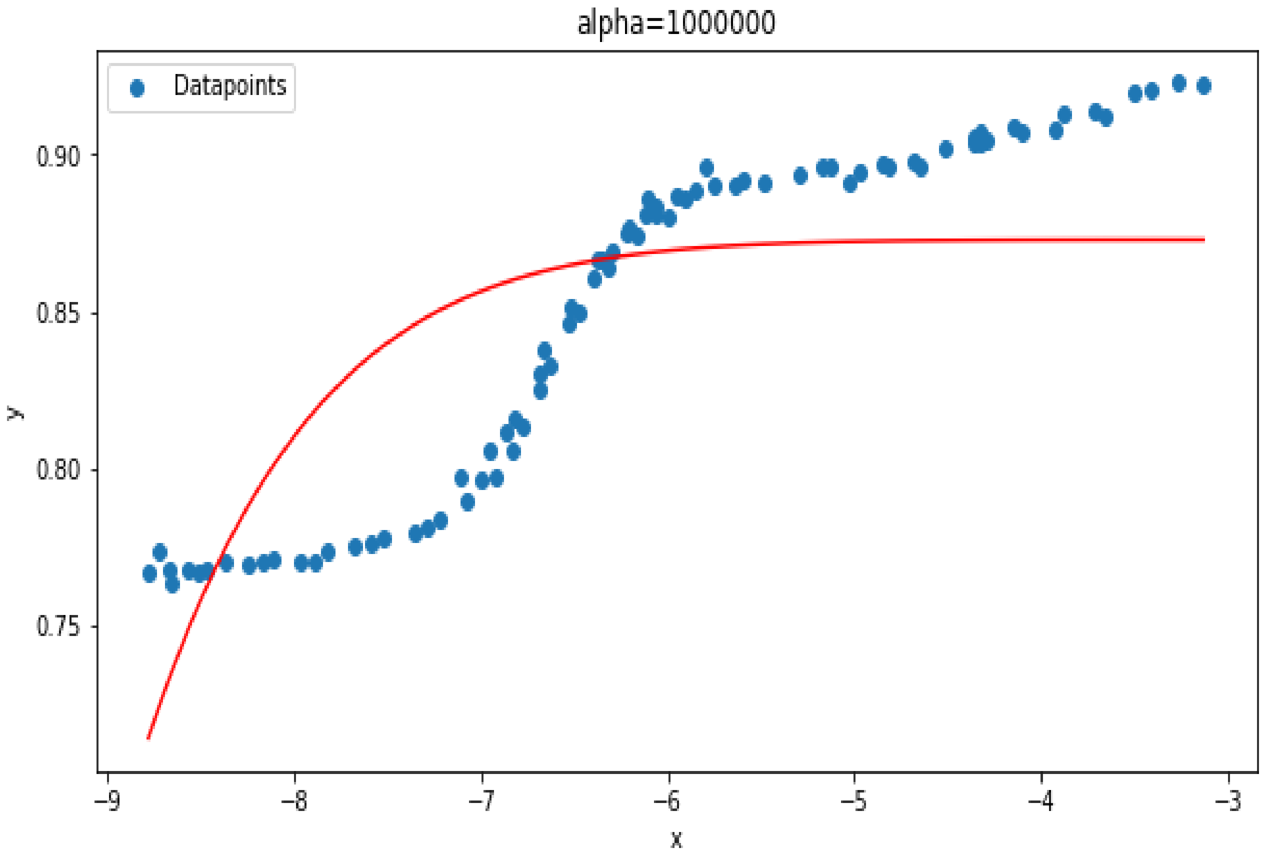
output when alpha = 0.5

c)



output when alpha = 10

d)



output when alpha = 1000000

c) Observation:

A large number of observations (denoted *n*) where *n* is large enough that the dataset can handle complex models. This is a hypothesis-generating approach to data analysis rather than a hypothesis-testing approach where statistical methods are used to determine the most predictive variables and build the model Penalized regression, especially the LASSO, can assist.

d) Conclusion:

The lasso could run completely in parallel. Using cross-validation, the lasso provides a validation of the covariate model. When it becomes too large, the algorithm starts modelling intricate relations to estimate the output and ends up overfitting the particular training data. The Lasso class takes in a parameter called alpha which represents the strength of the regularization term. A higher alpha value results in a stronger penalty, and therefore fewer features being used in the model. In other words, a higher alpha value such as 10 results in more features being removed from the model than a value such as 0.5.