1) Observation from the data:-

Feature 1 :- Initiall observation tells that linear regression will best fit feature 1 in Fig1.
Increasing to higher degree polynomial might overfit the training data and will give wrong prediction on test or cross validation data (CV).

— 0.5 marks

Feature 2 :-

From Fig 2 it can be observed that for feature 2 vs target plot 2 or 3 degree polynomial regression will give the best fit line.

— 0.5 marks

Feature 3 :- In Fig 3, we have multiple target values for 1 value of feature3.
Initiall observation tells that polynomial regression with higher degree around 4/5 might best fit feature 3.

— 0.5 marks

1) (a) & (b) has been tabulated in table 1.

→ 0·5 for table

→ 0·5 for comment

~~Justification:~~

Feature 1: (Fig 4)

Polynomial Regression with degree 1 i.e linear regression best fits the data. This can be seen from Fig 4 (b). Increasing to higher degrees (≥ 2) overfits the data and trys to capture the outliers.

This overfitting of training data leads to high error in cv/test data.

— 0·5 marks

Feature 2: (Fig 5)

Polynomial regression with degree 2 best fits the data (Fig 5c). Beyond degree 2 there is overfitting. Hence, we choose degree 2 polynomial regression for feature 2.

— 0·5 marks

Feature 3: (Fig 6)

As we have predicted from the initial observation that higher degree polynomial regression of around 4/5 might best fit the data.

Hence, ~~we plotted~~ multiple degrees of polynomial regression is performed. From Fig 6 we can see that from degree 4 we tend to have similar best fit line. Therefore, choosing degree less than 4 will be underfit and above 4 ~~will be~~ all give similar fit. Higher degrees of regression will be computationally extensive too. Hence, degree 4 is the best fit line for feature 3.

— 0.5 marks

Feature 1 & Feature 2

The data is mainly located at the center. The plane with degree 1 i.e bivariate linear regression ~~with~~ will best fit the data

— 1/3 marks.

Table — 1/3 marks

Comment — 1/3 marks

Feature 2 & 3

Acc to the initial observation,
The data is spread such that a plane with degree 1 or atmost degree 2 with best fit the data. This will be validated by the compution method.

— 1/3 marks.

Table — 1/3 marks

Comment — 1/3 marks.

Feature 1 & 3

Initial observation tells that ~~a~~ line will best fit the data.

Data looks like it is surrounded around a line like a noise.

Table — 1/3 marks

Comment — 1/3 marks.

Plot of best fit plane/ Line/ surface for Feature 1 & 2 — $\frac{1}{3}$ marks

Plot of best fit plane/ line/ surface for Feature 2 & 3 — $\frac{1}{3}$ marks

Plot of best fit plane/ line/ surface for Feature 1 & 3 — $\frac{1}{3}$ marks

~~3) Table (3) — 1 marks.~~

~~Comment — 1 marks.~~

3) Refer Table 3 — 1 marks.

Comment — 1 marks.

In multivariate linear regression, all the features are taken and regression analysis is performed. Multivariate regression gives better performance when compared with linear and bivariate. This is because in this we predict the target value based on multiple dependent features. Hence, we get better fit over the target..

Solution 1:

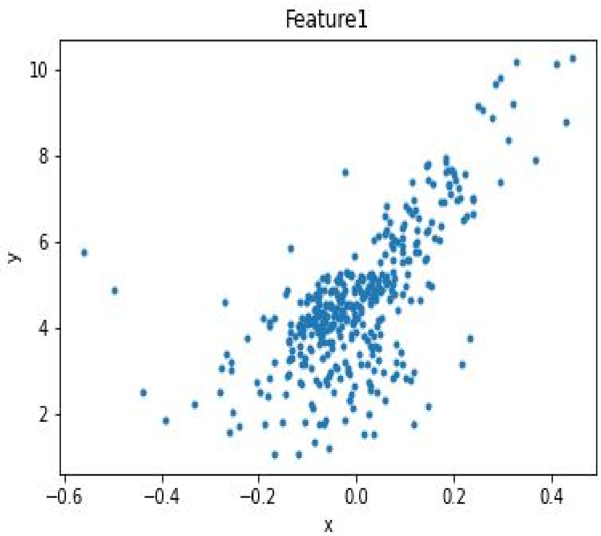


Fig 1 : Feature1 vs target

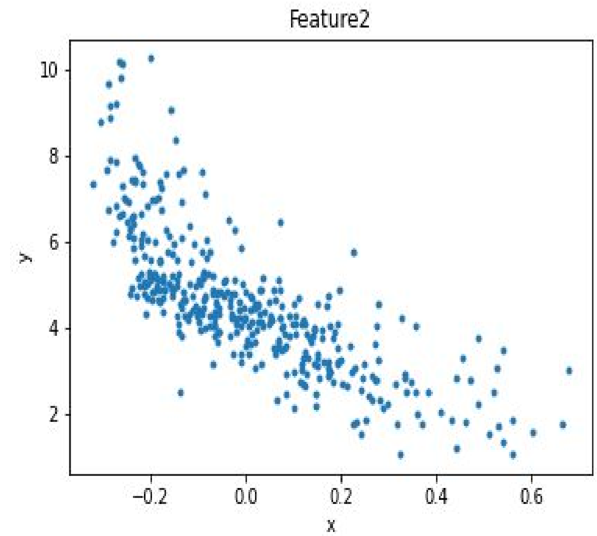


Fig 2 : Feature2 vs target

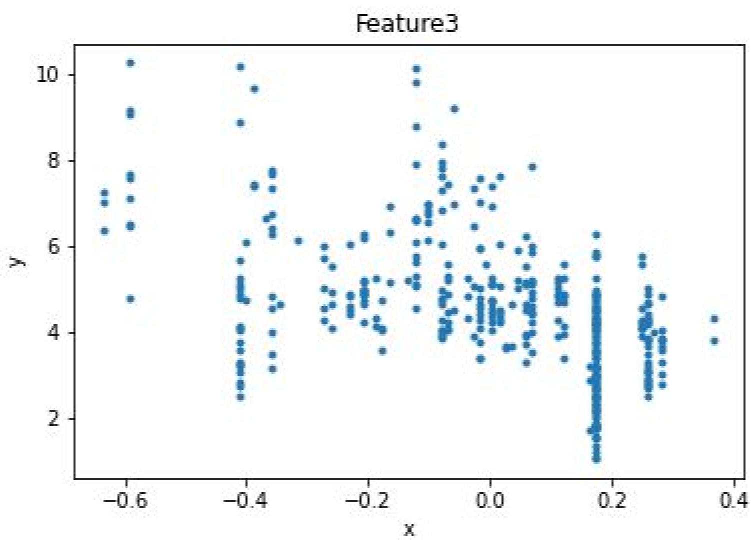


Fig 3 : Feature3 vs target

Features	k-fold	Parameters		Error		Comment
		w0	w1	Training Error	CV Error	
Feature1	Dataset 1	4.49	8.25	1.41	1.428	All the 5-fold gives approaximately same parameters. This shows that best fit line fits properly for all set from the shuffled dataset. From the data set:- When training error is high CV error is low and vice versa. Out aim is to pick the model which neither overfits nor underfits the data. Hence, we pick the model in which training error is closest to CV error.
	Dataset 2	4.494	8.626	1.424	1.432	
	Dataset 3	4.622	8.643	1.431	1.528	
	Dataset 4	4.51	8.373	1.264	1.998	
	Dataset 5	4.538	8.472	1.499	1.3379	
Feature2	Dataset 1	4.496	-6.24	1.071	1.516	
	Dataset 2	4.49	-6.046	1.115	1.263	
	Dataset 3	4.622	-6.358	1.235	1.118	
	Dataset 4	4.513	-6.369	1.093	1.406	
	Dataset 5	4.538	-6.212	1.233	0.754	
Feature3	Dataset 1	4.49	-3.62	1.96	2.146	
	Dataset 2	4.494	-3.8	2.023	1.917	
	Dataset 3	4.622	-3.888	1.95	2.499	
	Dataset 4	4.513	-3.398	1.963	2.129	
	Dataset 5	4.538	-4.016	2.068	1.7611	

Table1 : Univariate linear regression

Feature 1:-

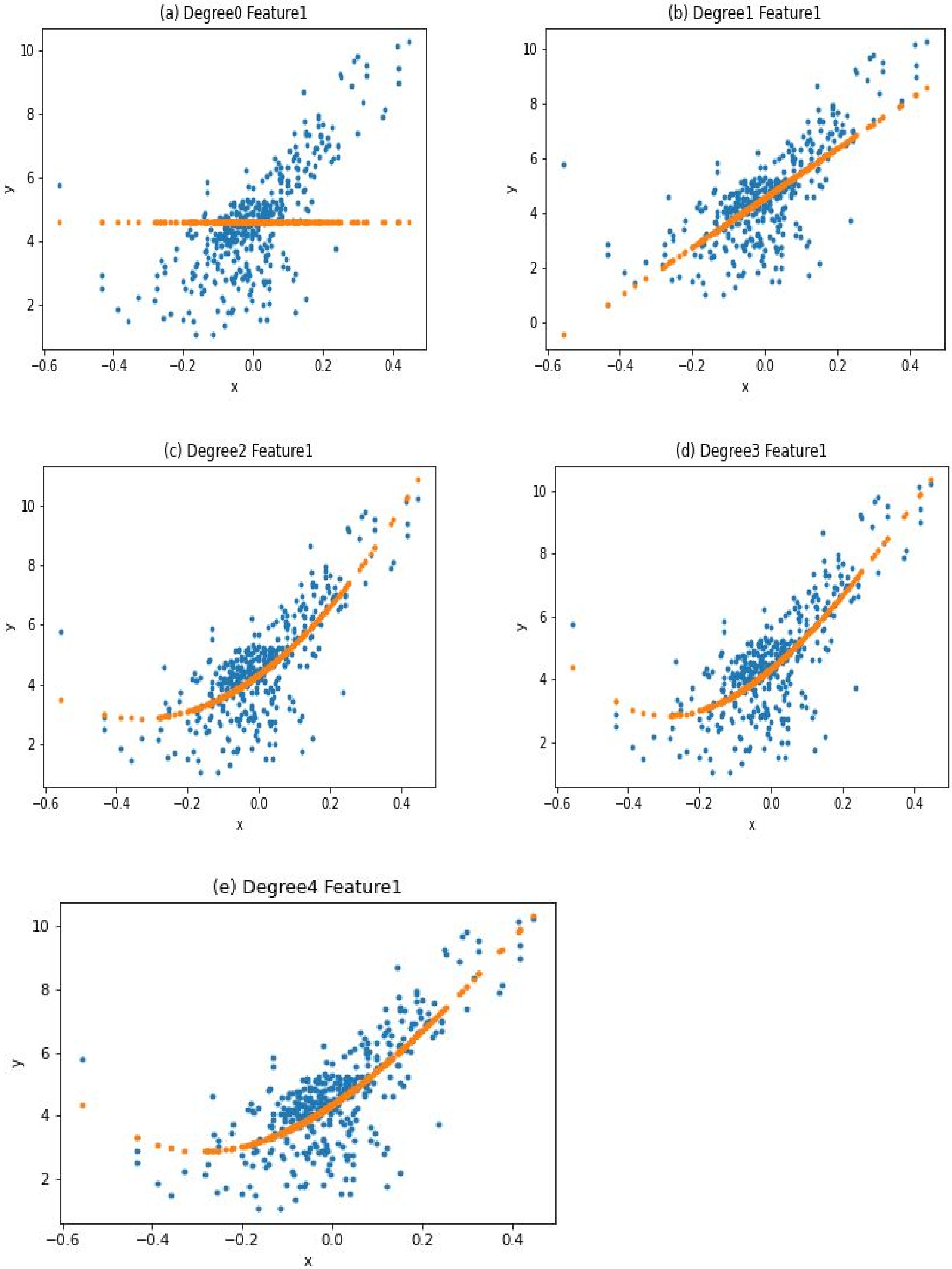


Fig 4 : Feature 1 univariate linear regression

Feature 2:-

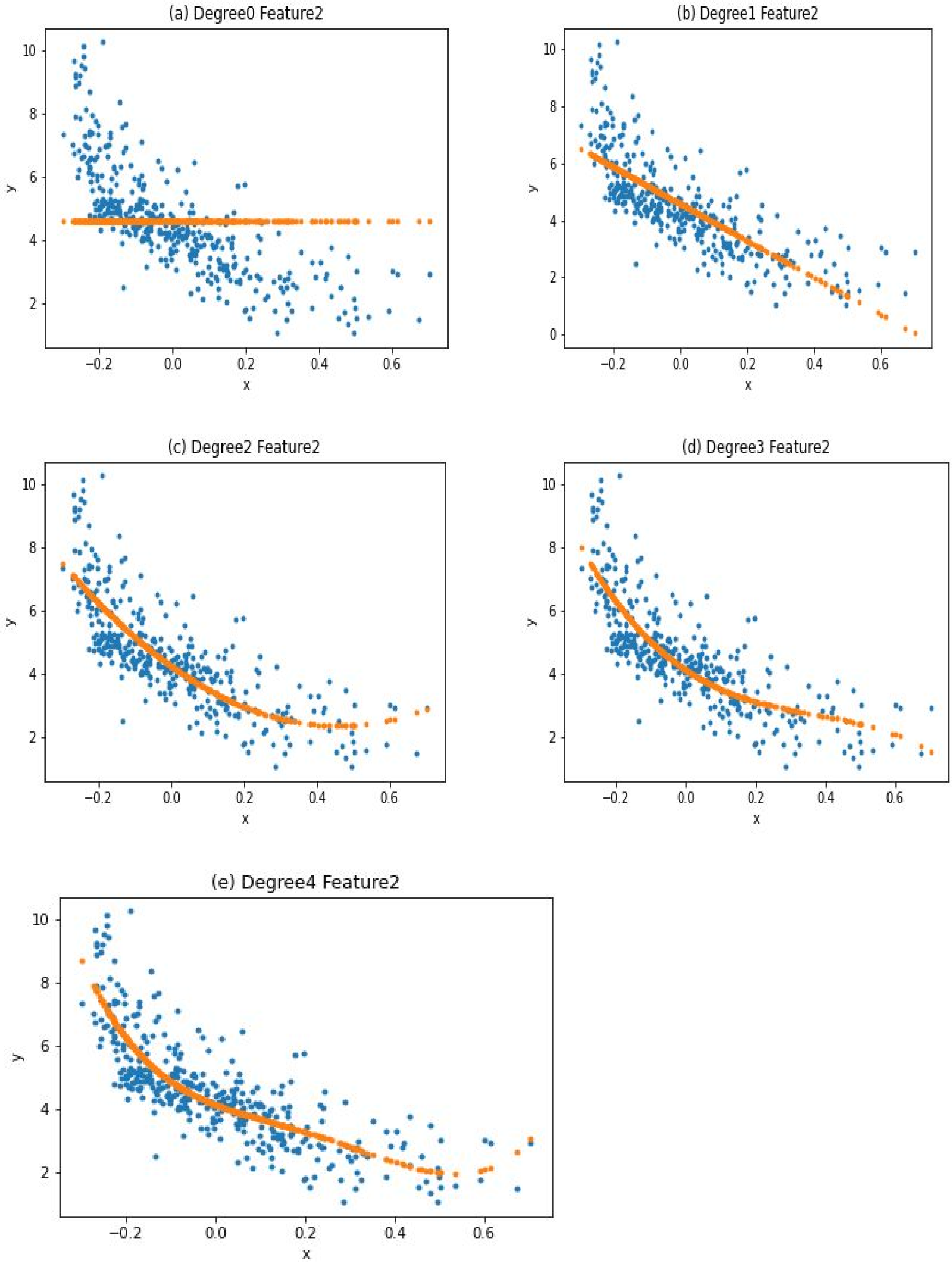


Fig 5 : Feature 2 univariate linear regression

Feature 3:-

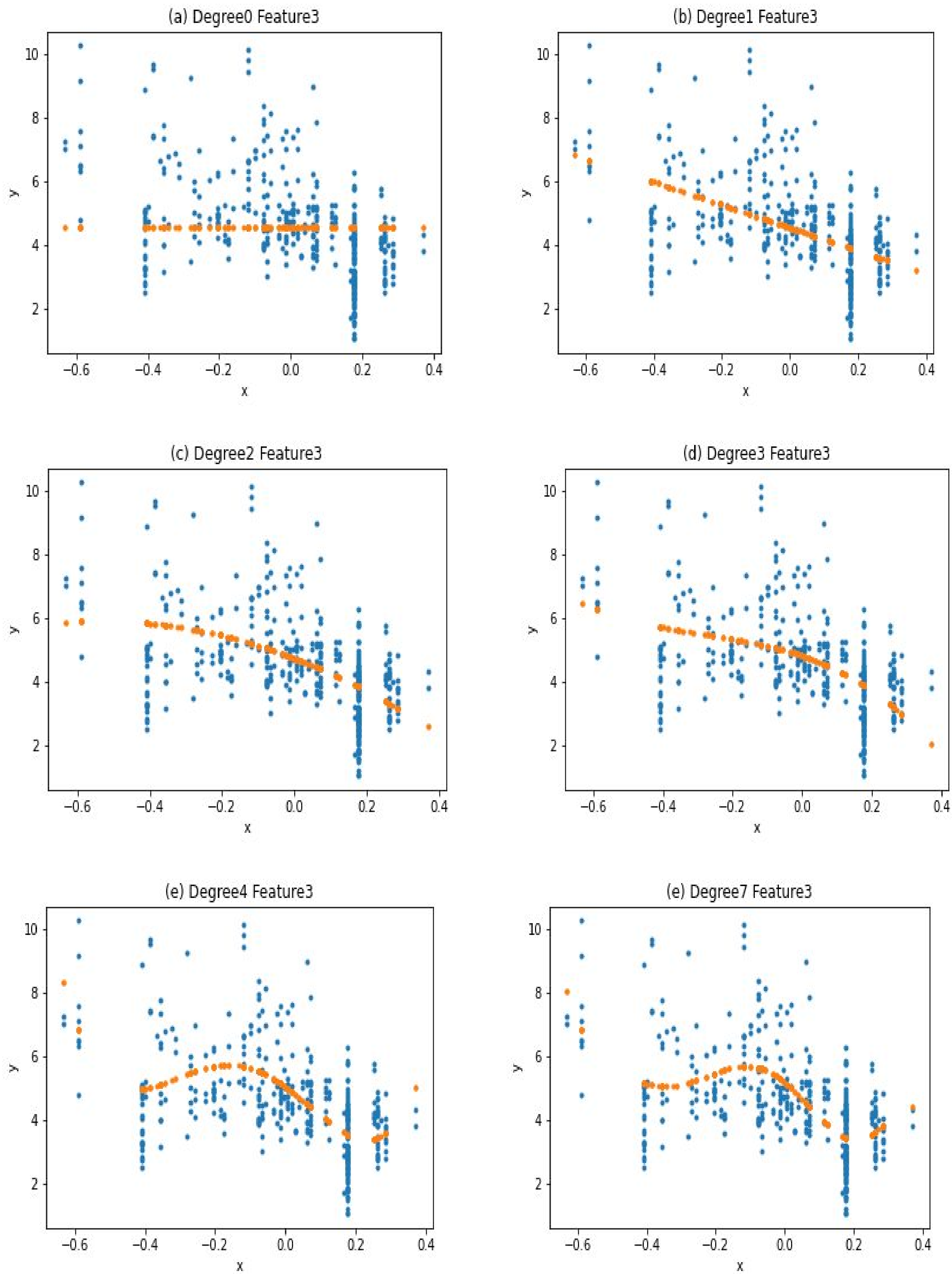


Fig 6 : Feature 3 univariate linear regression.

Solution 2:-

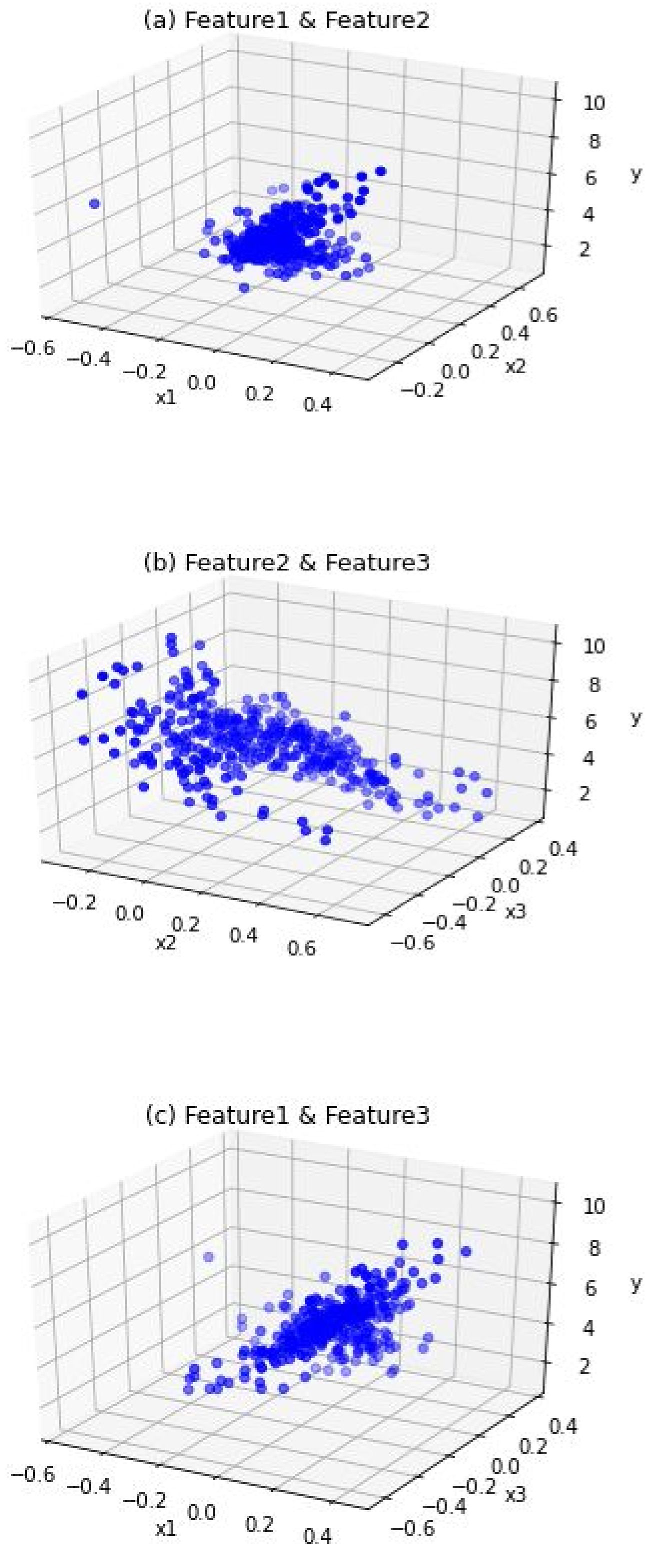
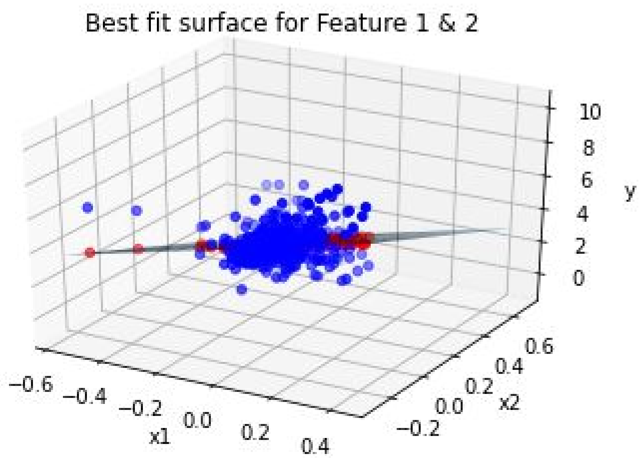
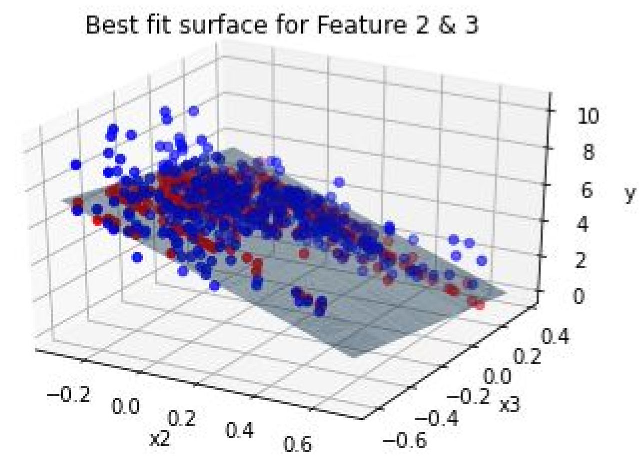


Fig 8 : Bivariate features vs target

In the figure 9 red are predicted values and blue are actual value





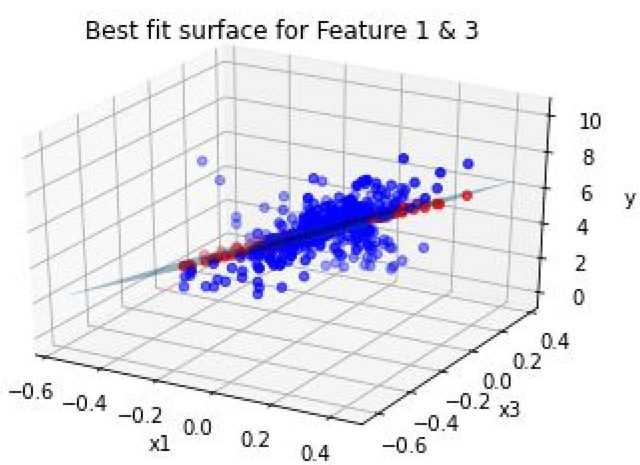


Fig 9(*a*), (*b*), (*c*) *contain actual value* (*blue*), *predicted value* (*red*), *best fit plane*.

Features	k-fold	Parameters			Error		Comment
		w0	w1	w2	Training Error	CV Error	
Feature(1&2)	Dataset 1	4.538	4.441	-4.358	0.997	0.5525	All the 5-fold gives approaximately same parameters. This shows that best plane fits closely for all set from the shuffled dataset. From the data set:- When training error is high CV error is low and vice versa. Out aim is to pick the model which neither overfits nor underfits the data. Hence, we pick the model in which training error is closest to CV error for all the features.
	Dataset 2	4.494	4.652	-4.354	0.85	1.224	
	Dataset 3	4.622	4.826	-4.3741	0.97	1.019	
	Dataset 4	4.5135	4.6077	-4.294	0.844	1.218	
	Dataset 5	4.538	7.127	-2.637	1.178	1.119	
Feature(2&3)	Dataset 1	4.496	-5.498	-1.929	0.91	1.22	
	Dataset 2	4.494	-5.276	-1.962	0.948	1.003	
	Dataset 3	4.622	-5.391	-2.198	1.016	1.09	
	Dataset 4	4.513	-5.591	-1.766	0.938	1.073	
	Dataset 5	4.438	-5.34	-2.352	0.982	0.911	
Feature(1&3)	Dataset 1	4.496	7.089	-2.399	1.149	1.043	
	Dataset 2	4.494	7.391	-2.518	1.132	1.206	
	Dataset 3	4.622	7.205	-2.492	1.141	1.285	
	Dataset 4	4.513	7.322	-2..218	1.006	1.615	
	Dataset 5	4.538	7.127	-2.637	1.178	1.119	

Table 2 : Bivariate linear regression

Solution 3:-

Features	k-fold	Parameters				Error	
		w0	w1	w2	w3	Training Error	CV Error
Feature(1, 2 & 3)	Dataset 1	4.45	3.68	-3.96	-1.91	0.72	1.07
	Dataset 2	4.56	3.59	-4.06	-1.78	0.78	0.67
	Dataset 3	4.58	4.55	-3.83	-1.74	0.78	1.09
	Dataset 4	4.55	4.36	-3.96	-1.86	0.77	0.81
	Dataset 5	4.54	4.32	-3.64	-1.95	0.73	0.92

Table 3 : Multivariate linear regression