1.

Table 1- Input data

ACTIVITIES	PROBABILITY	CUMULATIVE PROBABILITY
MOVIES	0.2	0.2
INF8245E	0.4	0.6
PLAYING	0.1	0.7
STUDYING	0.3	1

1-1

```
function randomActivity() {
        r ← generate U(0,1)
        if r<0.2 {
                activity = "movies"
        } if r<0.6 {
                activity = "inf8245e"
        } if r<0.7 {
                activity = "playing"
        } else {
                activity = "studying"
        }
        return activity
}
```

1-2

Table 2 - Replication results

REPLICATIONS	MOVIES	INF8245E	PLAYING	STUDYING	$\sum \lvert \bar{y}_i - p_i \rvert$
100	0.19	0.45	0.1	0.26	**0.1**
1000	0.21	0.39	0.09	0.31	**0.046**
P	0.2	0.4	0.1	0.3	

For comparing these two outputs, we can use the sum of absolute differences between probabilities and fractions. As it is shown in Table 2, this measure is approximately half of 100 replications for 1000 replications. Therefore, the more number of replications, results the less differences between probabilities and ration of activities.

2-1-(a) and (b)

The RMSE of Train Dataset: 2.5446

The RMSE of Valid Dataset: 37.6148

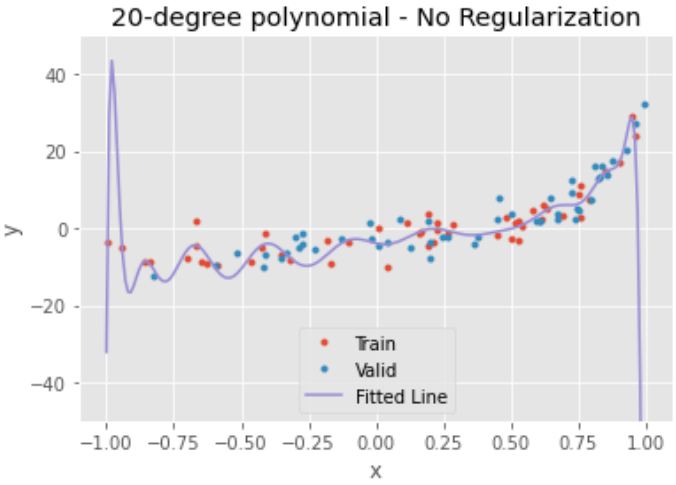


Figure 1 Model fitted on data

c)

It is involved with overfitting, because two reasons:

- As it is shown in Figure 1, the model is not supposed to follow the data before and after our dataset and it is overfitted on some data points in the train dataset.
- W values increase when the power of the variables increases.

2-2-(a) and (b)

Minimum of RMSE for train dataset: 2.984 , Corresponds to λ = 0.01

Minimum of RMSE for valid dataset: 3.043 , Corresponds to λ = 0.01

Minimum of RMSE for test dataset: 3.305 , Corresponds to λ = 0.01

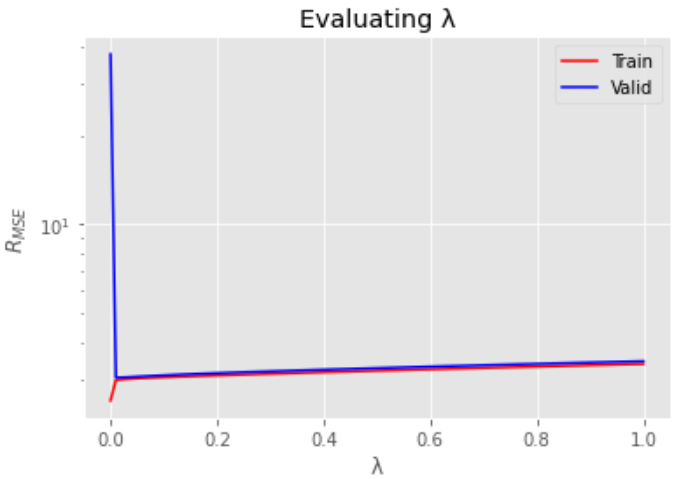


Figure 2 Evaluating λ

It seems from Figure 2 that the boundary for evaluating λ is not properly chosen. In this section, the R_{MSE} calculated for a λ in the logarithmic boundary of 10^{-6} and 1.

Minimum of RMSE for train dataset: 2.988 , Corresponds to λ = 0.0126

Minimum of RMSE for valid dataset: 3.042 , Corresponds to λ = 0.0126

Minimum of RMSE for test dataset: 3.302 , Corresponds to λ = 0.0126

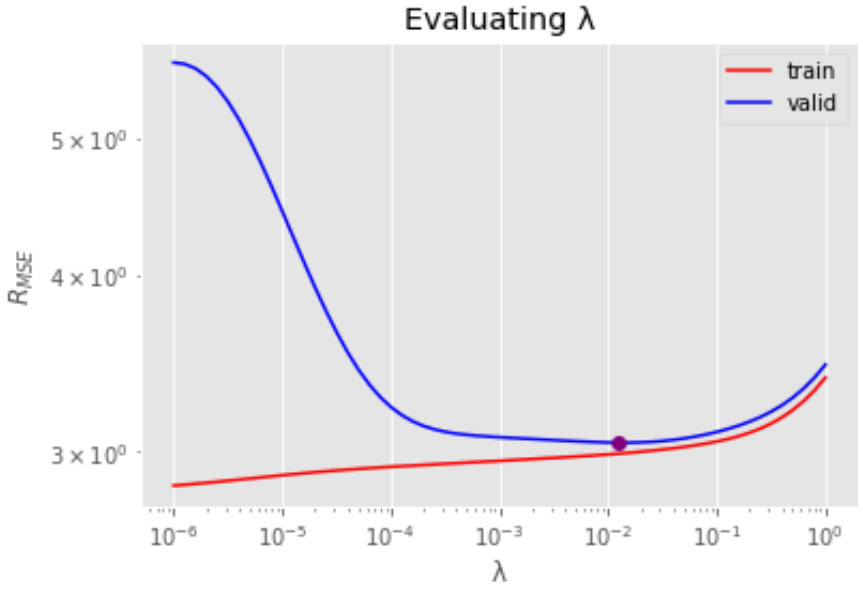
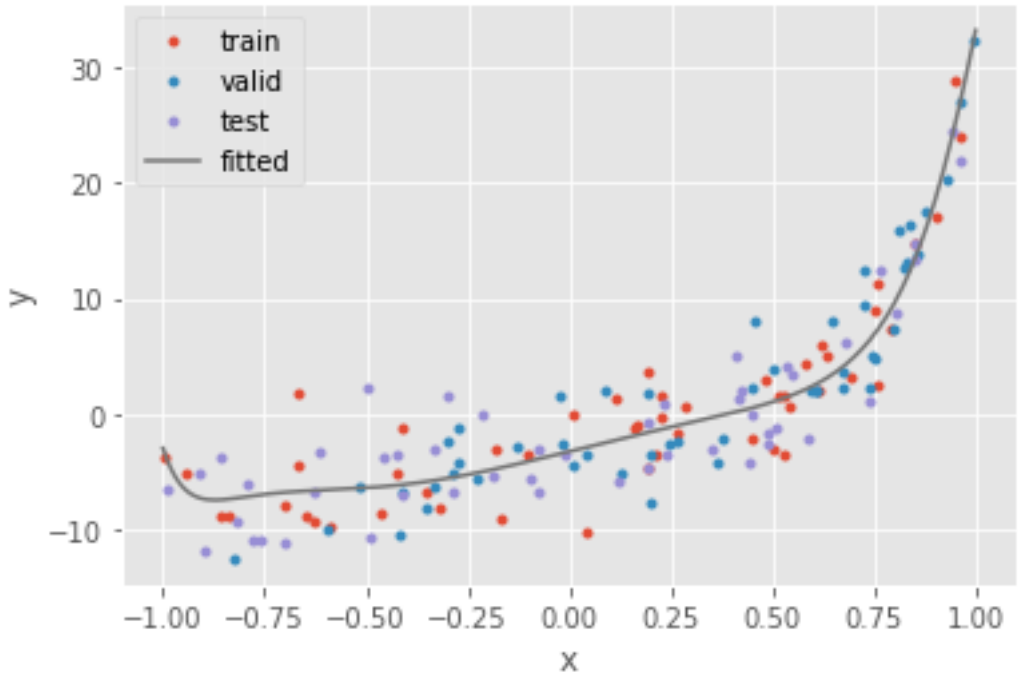


Figure 3- Evaluating λ

(c)



(d)

It is not supposed that the model is involved with overfitting. However, it is better to check its behavior near -1, since its movement has changed suddenly.

2-3

In the above figure, it seem the number of polynomial is an odd number since it has a lowering and an increasing. With respect to the length of elongation in between, I can estimate it is a 4 or 6 degree polynomial.

3-1-(a)

Minimum of RMSE for train dataset: 0.437 , Corresponds to N_epoch = 4459

Minimum of RMSE for valid dataset: 0.385 , Corresponds to N_epoch = 4459

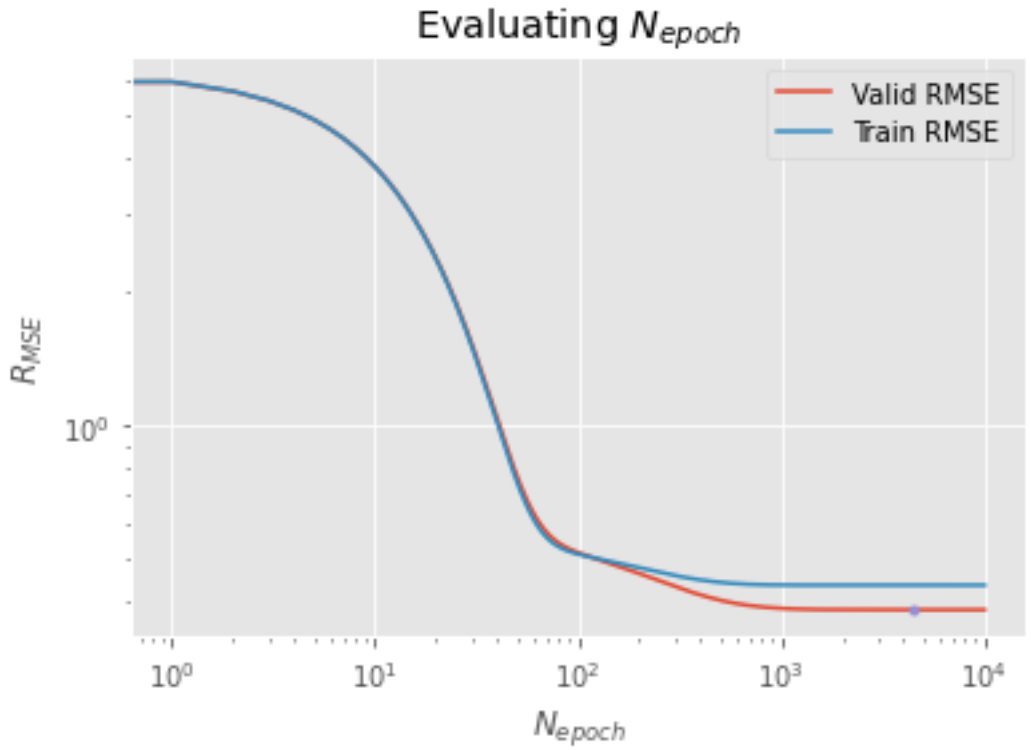


Figure 4- Evaluating N_{epoch}

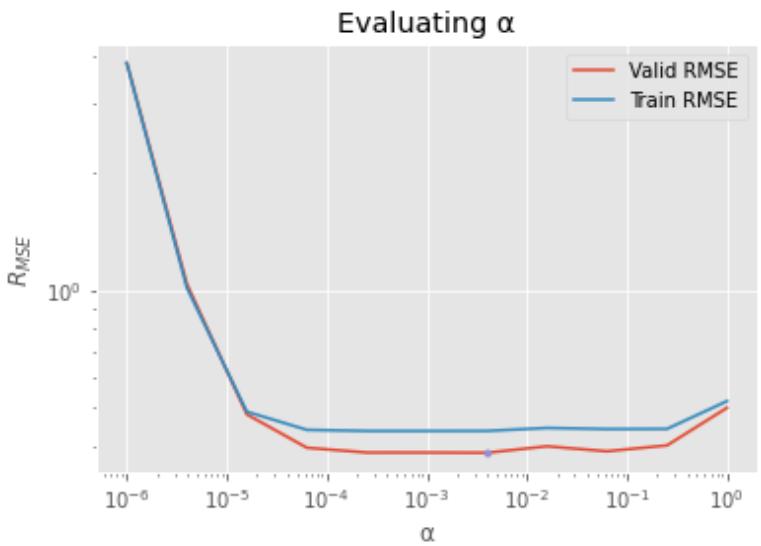
It seems it converges at 1000 number of epochs. Therefore, I set the N_{epoch} = 1000.

3-2

MINIMUM OF RMSE FOR TRAIN DATASET: 0.437, CORRESPONDS TO A = 0.004

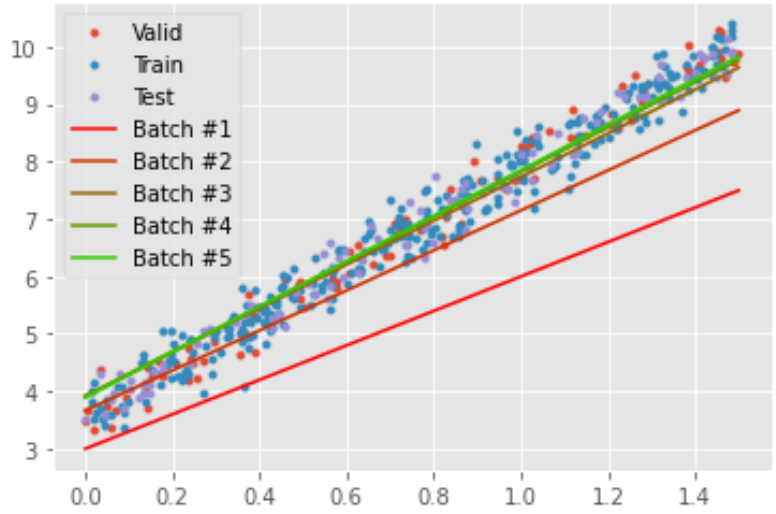
MINIMUM OF RMSE FOR VALID DATASET: 0.384, CORRESPONDS TO A = 0.004

MINIMUM OF RMSE FOR TEST DATASET: 0.372, CORRESPONDS TO A = 0.004



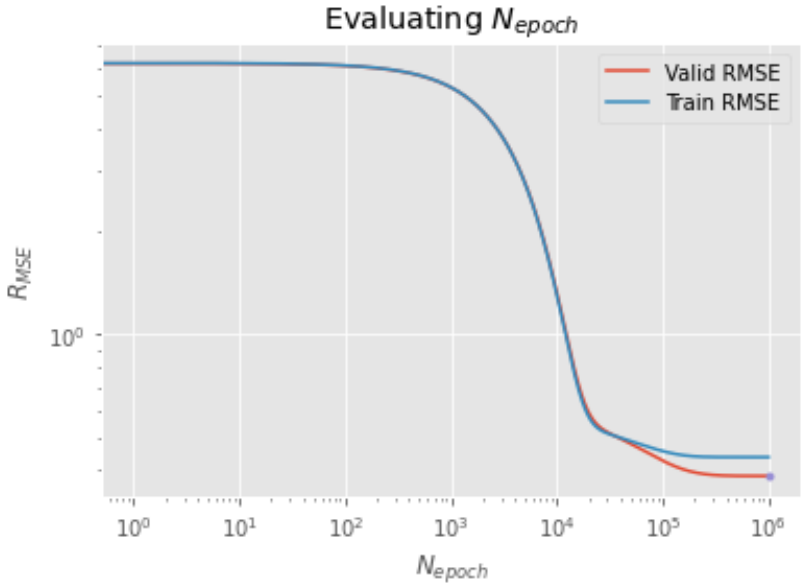
α	$R_{MSE\|valid}$
0.000001	3.824
0.000004	1.040
0.000016	0.482
0.000063	0.396
0.000251	0.385
0.001000	0.385
0.003981	0.384
0.015849	0.399
0.063096	0.388
0.251189	0.401
1.000000	0.501

3-3



3-4

Minimum of RMSE for valid dataset: 0.385 , Corresponds to N_epoch = 999999



It approximately converges for more than 50,000 epoches.

3-5

The difference between the number of epochs in two approaches are extensive. A reason for this issue is the learning rate, which was equal in both approaches. In every epoch in the first approach, N times a gradient learns the W parameter. However, in the second approach, the whole dataset learns the W for one time. Therefore, the value of alpha should be higher in the ridge regression since it has more reliable direction to the minimum. Also, it may be because stocking gradients descend in the local minimum.

4-1

In the mean approach, it may change the distribution of data in result of reducing the standard deviation. Therefore, we can generate a random number with the normal distribution, in order to remaining the standard deviation. However, generated random number maybe a non-positive value. In continue, I will use the Mean approach to fill the data.

	POLICOPERBUDG	COLUMN	
#	**sample**	**Filled with mean**	**Filled with normal**
0	0.04	0.040000	0.040000
1	NaN	0.076708	0.192581
2	NaN	0.076708	0.067859
3	NaN	0.076708	0.045014
4	NaN	0.076708	-0.038052

4-2

RMSE of validation set (fold 1)= 0.166

RMSE of validation set (fold 2)= 0.137

RMSE of validation set (fold 3)= 0.132

RMSE of validation set (fold 4)= 1.082

RMSE of validation set (fold 5)= 0.137

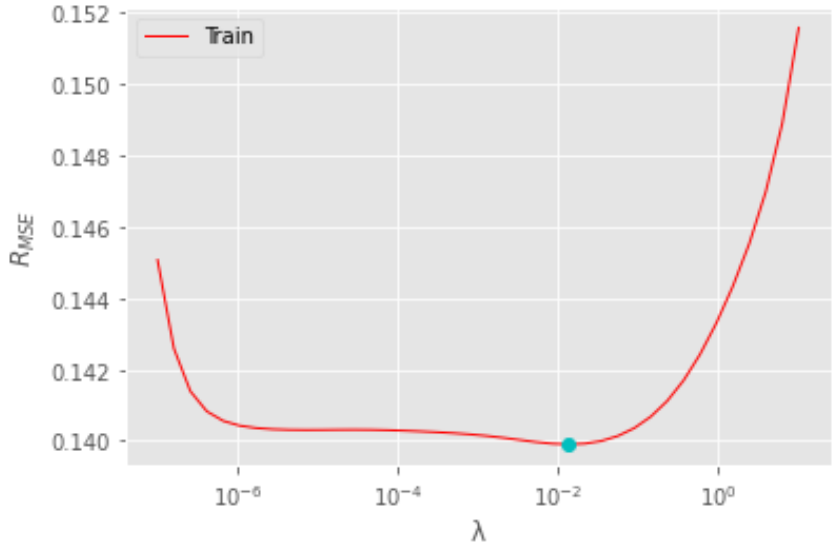
Average RMSE for 5-fold Train dataset = 0.331

RMSE for Test dataset = 0.142

4-3-(a), (b) and (c)

Minimum of RMSE for train dataset: 0.140 , Corresponds to λ = 0.0160

RMSE for test dataset: 0.143 , Corresponds to λ = 0.0160



I use λ in a logarithmic scale from 10^{-7} and 10^{0}. In this way, we can observe the parameter in a wide logarithmic range with more sensitivity near zero. ($\lambda > 0$)

4-3-(d)

Yes, the features with bigger β has more effect on the model. but the distribution of the feature is also important. To standardize the value we can calculate $\bar{\beta} = \beta/mu$. Features with more value of $\bar{\beta}$ can be a good descriptive candidate for our model. Therefore, I choose the N features with the biggest $\bar{\beta}$ for my model.

4-3-(e)

25 Important Feature as sorted:

> NumImmig, numbUrban, NumIlleg, LemasSwornFT, OfficAssgnDrugUnits, PctNotSpeakEnglWell, OwnOccLowQuart, whitePerCap, PctRecImmig5, PolicBudgPerPop, LemasTotalReq, PctPersOwnOccup, PctPopUnderPov, PctKids2Par, PctLargHouseOccup, PersPerRentOccHous, LemasSwFTPerPop, medIncome, RentLowQ, TotalPctDiv, agePct16t24, PersPerFam, PctPolicHisp, PctLess9thGrade, PctPolicBlack

Minimum of RMSE for train dataset: 0.146 , Corresponds to λ = 1e-07

RMSE for test dataset: 0.149 , Corresponds to λ = 1e-07

4-3-(f)

The R_{MSE} is a little higher than the model with full-features, which is rational since the number of features is less than the other model. However, this difference ($\cong 2.1\%$) is evaluated by the less number of features by 5 times (124 vs. 25) which causes using less computation and memory capacity as well as, less cost for data gathering.