Junyi CHEN

CS5487

2017/10/04

CS5487 Programming Assignment 1: regression

# PART 1 pOLYNOMINAL FUNCTION

## implementation

Implementation written in python is attached in the source code files.

## REGRESSION PLOTS AND HYPERPARAMETERS TUNNING

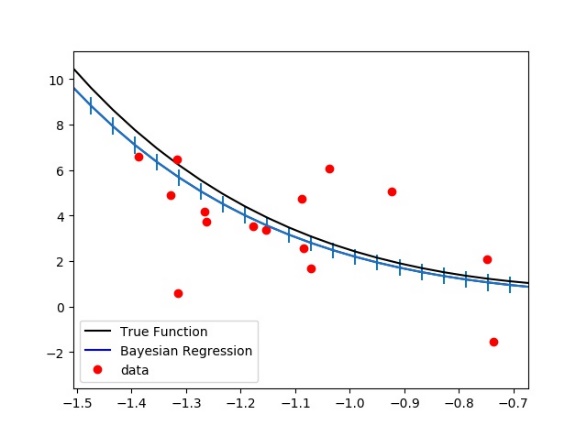
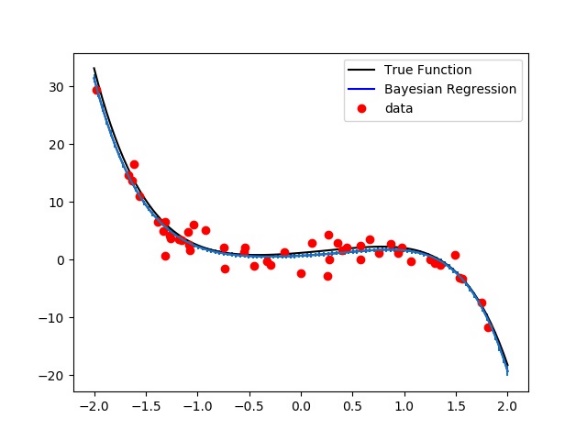
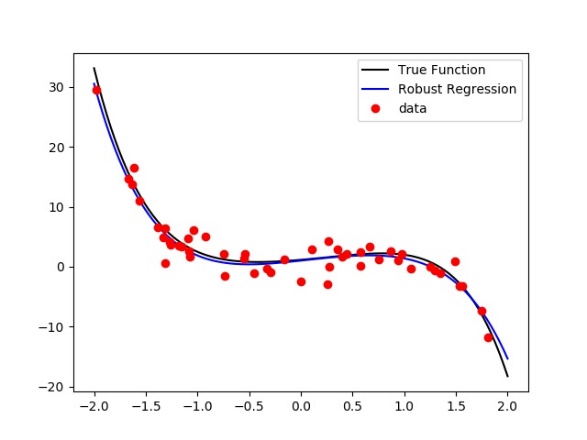
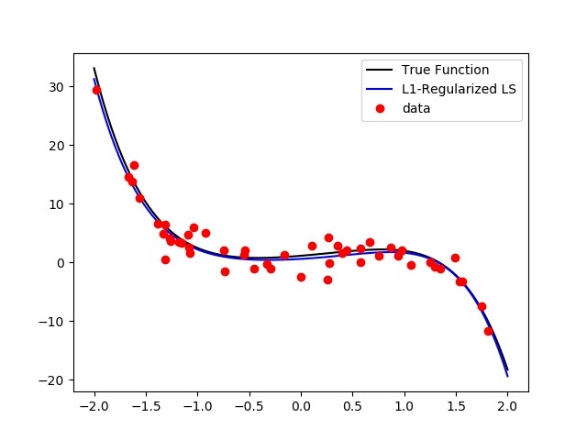
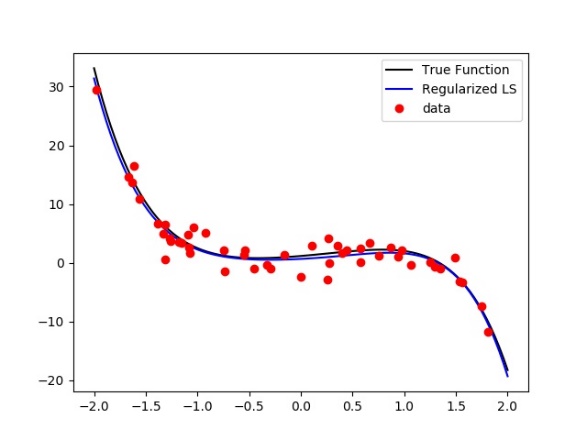
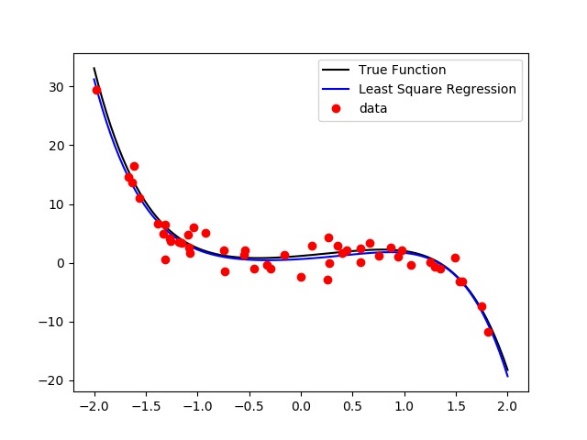


Figure 1: Predict function of different regression methods (figure in lower-right corner is the zoomed plot to show deviation of Bayesian Regression)

Figure 1 are plots of predictive functions of 5 different regression methods. Quadratic programming and linear programming solvers comes from the python package ‘cvxopt’. Hyper-parameters (if any) are chosen by examining mean-square errors shown in Table 1. Values in bold type corresponds to ‘optimal hyper-parameters’ in our experiments (noted that LS and RR required no hyper-parameters). From the table, we can see that RLS (Regularized LS) and BR (Bayesian Regression) have the smallest MSE around 0.4076, a bit smaller than LS (Least-squares) and LASSO (L1-regularized LS) with 0.4086 and 0.4128 respectively while RR (Robust Regression) has a relatively larger number around 0.7680.

Table 1: Experiment mean-square errors of different hyper-parameters

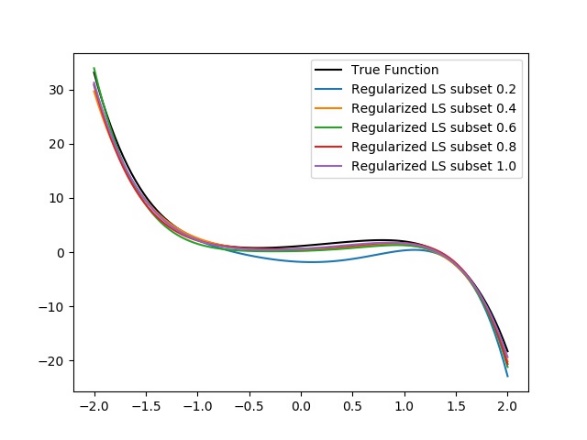
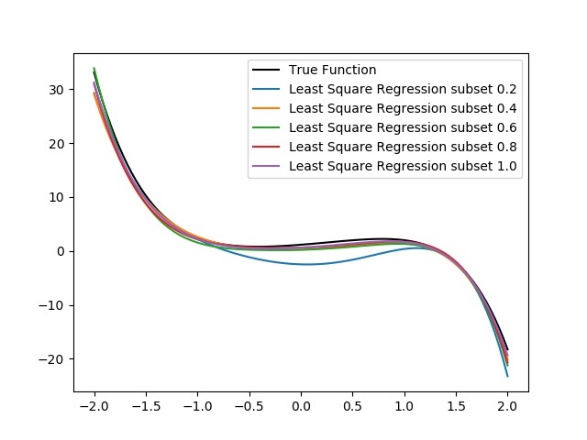
|  |  |  |
| --- | --- | --- |
|  | MSE of LS | MSE of RR |
| NA | 0.408644 | 0.768046 |

|  |  |
| --- | --- |
| Alpha and sigma | MSE of BR |
| 'alpha': 0.1, 'sigma': 0.1 | 0.408237 |
| 'alpha': 0.1, 'sigma': 0.5 | 0.420991 |
| 'alpha': 0.1, 'sigma': 1 | 0.557904 |
| 'alpha': 0.1, 'sigma': 5 | 2.895878 |
| 'alpha': 0.5, 'sigma': 0.1 | 0.408553 |
| 'alpha': 0.5, 'sigma': 0.5 | **0.4076** |
| 'alpha': 0.5, 'sigma': 1 | 0.415603 |
| 'alpha': 0.5, 'sigma': 5 | 1.238121 |
| 'alpha': 1, 'sigma': 0.1 | 0.408598 |
| 'alpha': 1, 'sigma': 0.5 | 0.407827 |
| 'alpha': 1, 'sigma': 1 | 0.408633 |
| 'alpha': 1, 'sigma': 5 | 0.856287 |
| 'alpha': 5, 'sigma': 0.1 | 0.408635 |
| 'alpha': 5, 'sigma': 0.5 | 0.408425 |
| 'alpha': 5, 'sigma': 1 | 0.407939 |
| 'alpha': 5, 'sigma': 5 | 0.459158 |

|  |  |  |
| --- | --- | --- |
| Lambda | MSE of RLS | MSE of LASSO |
| 0.1 | 0.408236518 | **0.412832** |
| 0.25 | 0.407826865 | 0.420168 |
| 0.5 | **0.407600224** | 0.434636 |
| 1 | 0.408632571 | 0.474607 |
| 2 | 0.415602545 | 0.519128 |
| 5 | 0.459157939 | 0.569842 |

## sample subsets and learning curves

In our experiment, subset sizes of samples are 20%, 40%, 60%, 80%. For each size of subset, we run 5 trials of different random subsets and take the average error. The sampling function comes from python package ‘scikit-learn’. On the other hand, we only plot the first-round function inside 10 same size trials. In this section, figure 2 include plots range from 20% to 80% of the full dataset using the hypermeters from the previous section, full set figure is not provided for it is identity to plots in part (b). Curves in different colors are prediction functions leanrned form different sizes of subset. Figure 3 are leraing curves indicate relations between mean square errors and size of subset. Range of y aixs is fixed, the part of curve with y value larger than 50 will not be displayed.



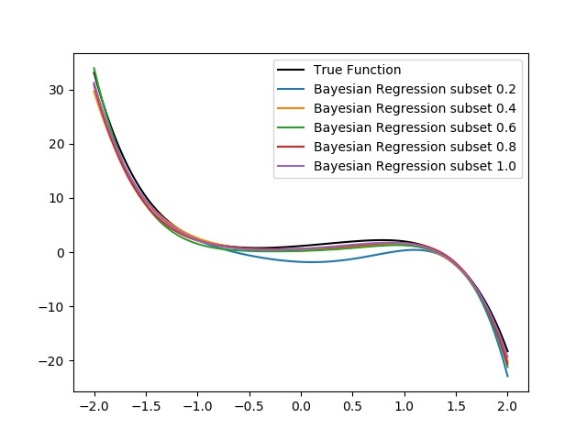
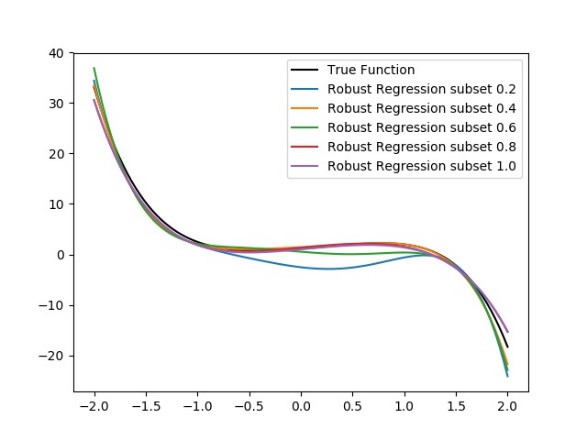
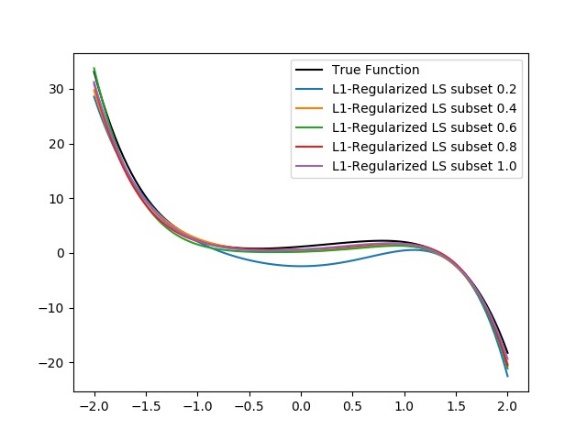
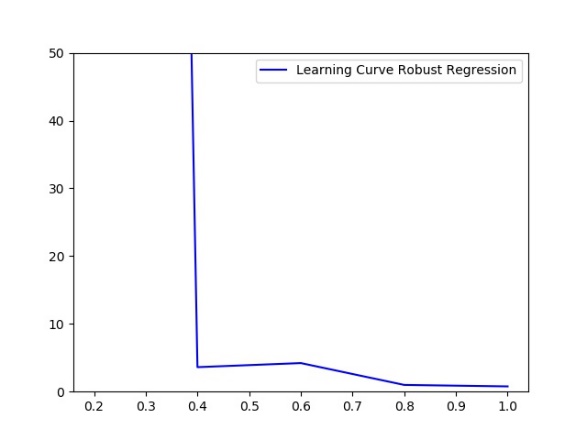
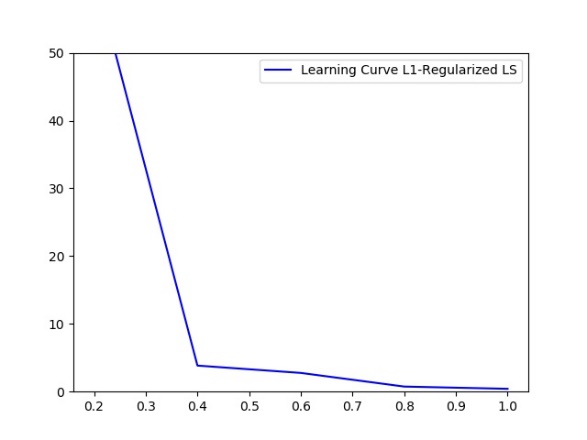
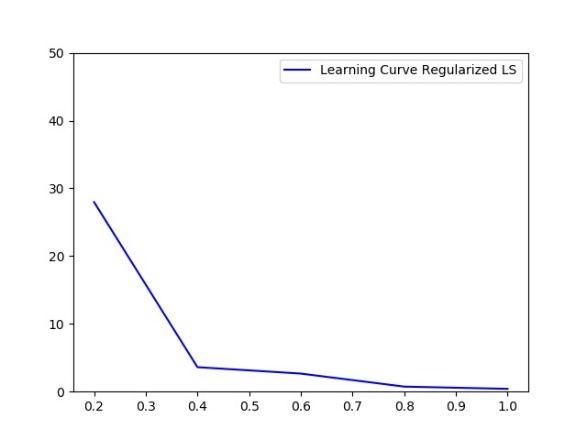
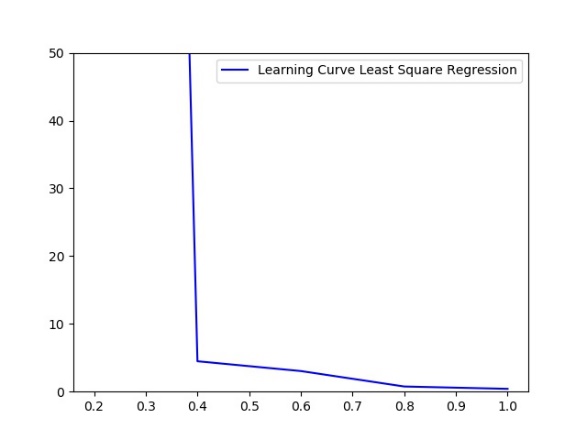


Figure 2 Prediction functions and true functions of different regression methods

From two set of figures, we can be inferred that RLS and BR have better regression performances when the dataset is small, for they have intuitively ‘closer’ line in the prediction function plots and much smaller MSE indicated by the learning curves. Besides, though LASSO have an MSE excesses the limit of 50 when we train the model using 20% of data, it is much better than LS and RR whose MSE are around 500 and 600 respectively. This experiment may indicate that RLS, LASSO and BR have better resistances against overfitting.

On the other hand, for RB regression, the mean square rise when the subset size increase from 40% to 60%, this phenomenon further implies that RB is vulnerable to overfitting.



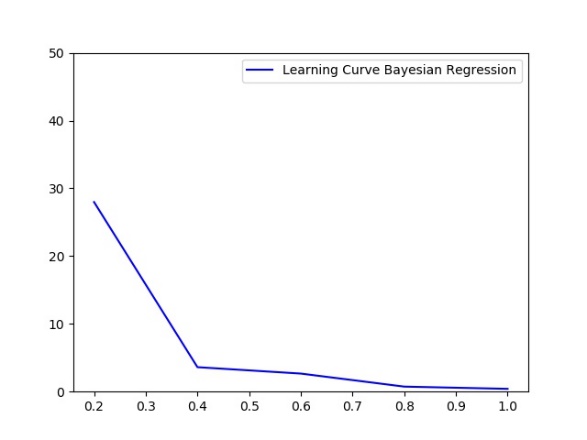


Figure 3 Learning curves of different regression methods

## outlies and robustity analysis

In this section, 4 outliers are added to the training set: (outliers\_x = [-1.3,0.5,0.7,1]; outliers\_y = [80,30,50,-30]). Hyper-parameters are set as we examined before. Table 2 and Figure 4 show some statistic of outliers experiments.

Table 2 MSE of different methods with outliers

|  |  |
| --- | --- |
|  | MSE |
| RR | **0.791648** |
| RLS | 6.943168 |
| BR | 6.943168 |
| LASSO | 9.413357 |
| LS | 9.508327 |

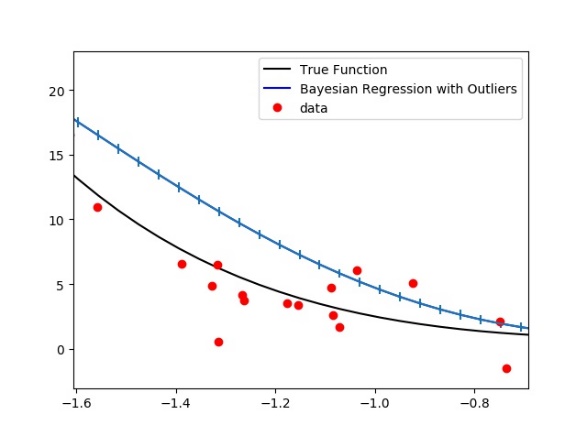
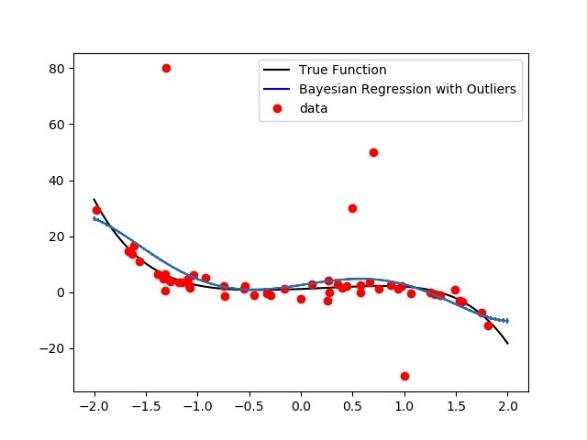
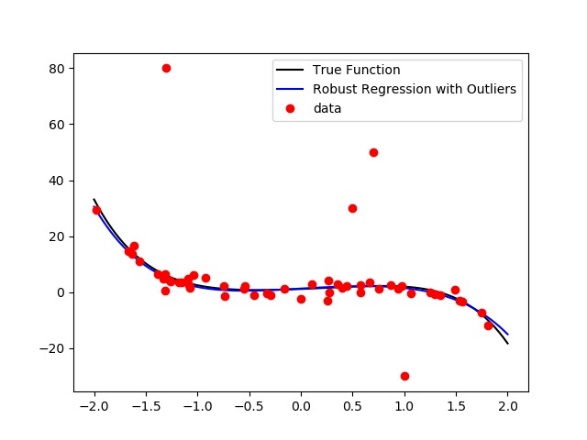
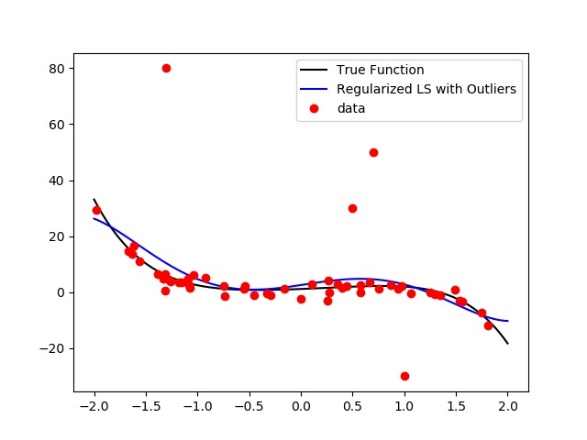
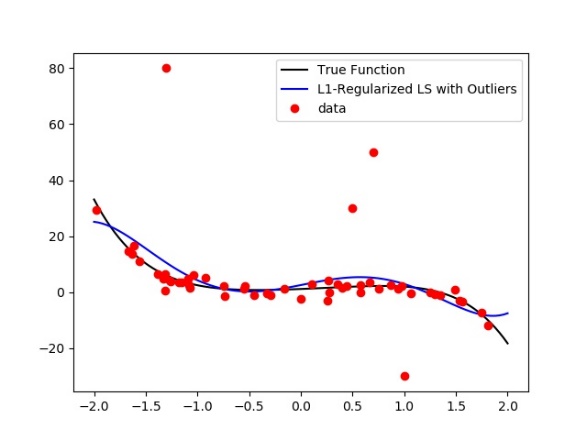
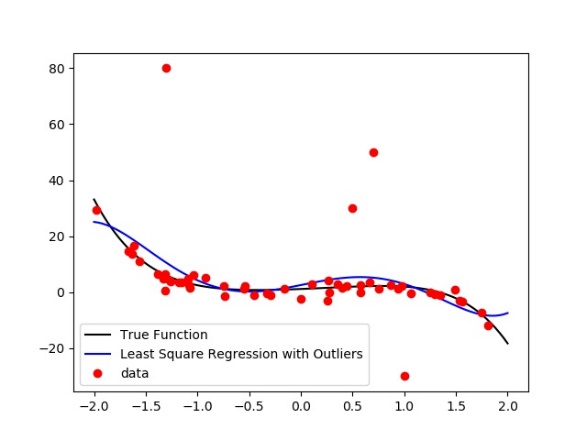


Figure 4 Predict function with outliers of different regression methods (figure in lower-right corner is the zoomed plot to show deviation of Bayesian Regression)

From figure 4, we can see that RR have the best performance to resist outliers for it has a similar curve with normal experiment. Also, MSE of RR stays at around 0.8 this is another evidence that RR is robust when some outliers are added to the dataset. RLS and BR have similar MSE around 7 while LASSO and LS have MSE around 9.5. In conclusion, RR is the robust to outliers LS is the most sensitive one.

One possible reason is that the objective function of LS is square formed. Outliers with unexpected large values will amplify the error function a lot so the regression method tend to be sensitive to the outliers.

## HIGHER ORDER ESTIMATION

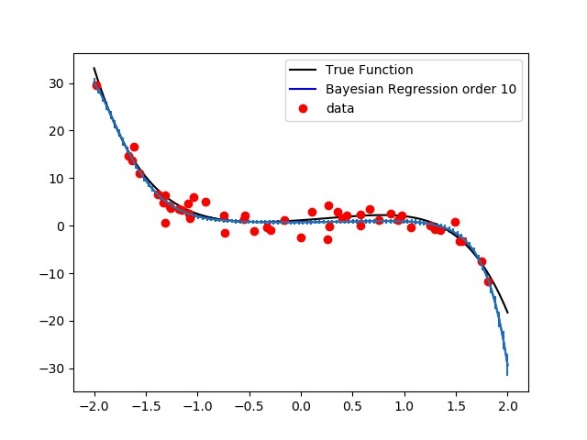
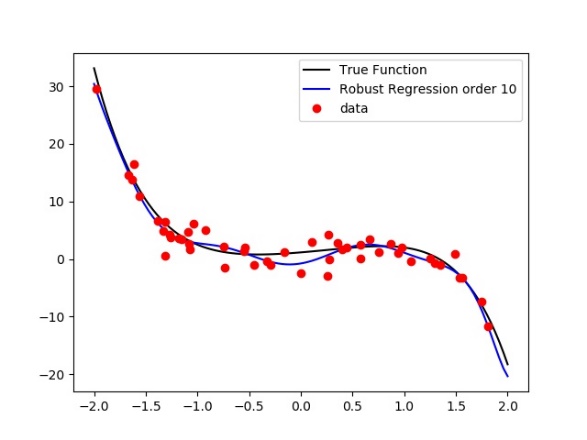
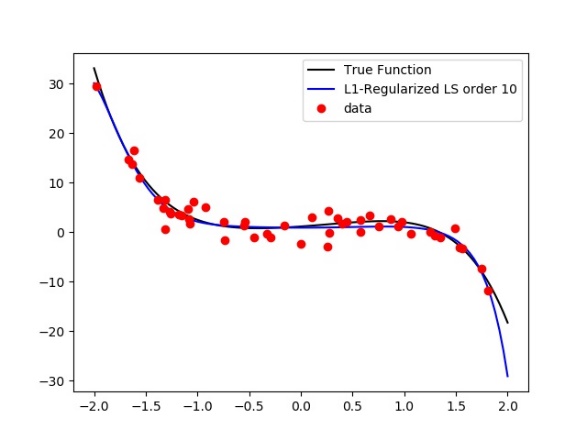
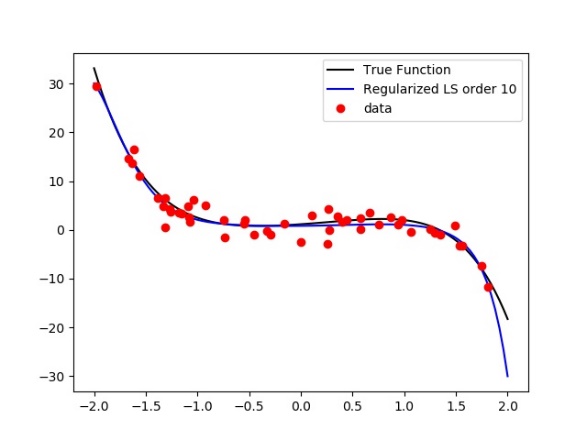
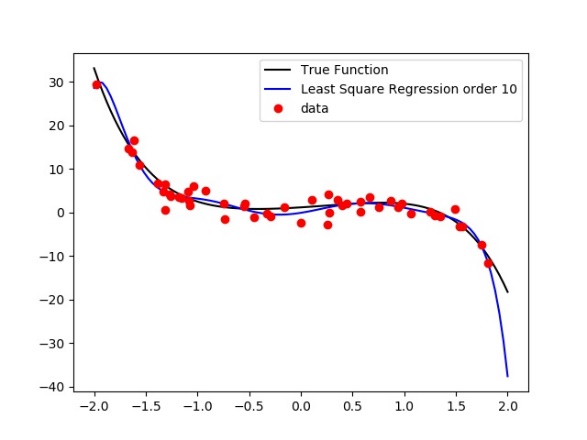


Figure 5 Predict function of different regression methods (order 10)

From Table 3, we can conclude that RR have the lowest MSE while LS outputs the largest MSE. RLS, LASSO and BR share similar numbers. But when we refer to Figure 5, the RR curve is twisted and not ‘close’ (visually) to the true function. The reason why it has a low MSE is that the curve is close to the true function on its two terminals when we scope the ‘extreme’ data points which contributes a lot to the MSE.

Table 4 is a list of predicted theta values of different. The yellow part of the table theta of order higher than 5. We can see that RR an LS tend to have large values in higher order scalars. Because we have knowledge that true function is a 5-order function. A well fitted set of thetas should have small theta when the order is higher than Phenomena on LS and RR, infer that two estimators overfit the dataset.

Table 3 Experiment mean-square errors of different hyper-parameters (order 10)

|  |  |  |
| --- | --- | --- |
|  | MSE of LS | MSE of RR |
| NA | 7.983107 | 1.289857 |

|  |  |
| --- | --- |
| Alpha and sigma | MSE of BR |
| 'alpha': 0.1, 'sigma': 0.1 | 17.82521114 |
| 'alpha': 0.1, 'sigma': 0.5 | 3.890454631 |
| 'alpha': 0.1, 'sigma': 1 | **2.876626237** |
| 'alpha': 0.1, 'sigma': 5 | 7.669525294 |
| 'alpha': 0.5, 'sigma': 0.1 | 15.91938565 |
| 'alpha': 0.5, 'sigma': 0.5 | 9.98381222 |
| 'alpha': 0.5, 'sigma': 1 | 4.351475446 |
| 'alpha': 0.5, 'sigma': 5 | 4.518810872 |
| 'alpha': 1, 'sigma': 0.1 | 13.49094405 |
| 'alpha': 1, 'sigma': 0.5 | 14.03001502 |
| 'alpha': 1, 'sigma': 1 | 6.548096165 |
| 'alpha': 1, 'sigma': 5 | 3.49799502 |
| 'alpha': 5, 'sigma': 0.1 | 9.573704116 |
| 'alpha': 5, 'sigma': 0.5 | 18.18799337 |
| 'alpha': 5, 'sigma': 1 | 15.22605129 |
| 'alpha': 5, 'sigma': 5 | 3.043253601 |

|  |  |  |
| --- | --- | --- |
| Lambda | MSE of RLS | MSE of LASSO |
| 0.1 | 17.82521 | 16.64525 |
| 0.25 | 14.03002 | 19.16953 |
| 0.5 | 9.983812 | 13.3456 |
| 1 | 6.548096 | 6.142715 |
| 2 | 4.351475 | **2.686592** |
| 5 | **3.043254** | 3.493389 |

Table 4 Theta values of different methods (some rows of LASSO is missing, possibly due to precision issues)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | LS | RLS | LASSO | RR | BR | True |
| 0 | -0.09938 | 0.819574 | 0.912094 | -0.77214 | 0.73481 | 1.15243 |
| 1 | 3.793512 | 0.219007 | 5.23E-08 | 3.025203 | 0.115234 | 1.48629 |
| 2 | 6.49122 | 0.601251 | 0.550202 | 13.14885 | 0.448381 | 0.92950 |
| 3 | -10.7448 | -0.29162 | NA | -7.57593 | -0.20169 | -1.11344 |
| 4 | -5.52168 | 0.067251 | NA | -19.78 | 0.13653 | 0.15980 |
| 5 | 8.632264 | -0.26496 | -0.35734 | 5.642387 | -0.23723 | -0.61788 |
| 6 | 0.653714 | -0.13463 | 2.48E-10 | 12.19446 | 0.00411 | 0 |
| 7 | -3.0458 | -0.19206 | -0.19844 | -2.1642 | -0.22735 | 0 |
| 8 | 0.763549 | 0.240247 | 0.171415 | -3.15724 | 0.15408 | 0 |
| 9 | 0.310281 | 0.00966 | 0.014124 | 0.245386 | 0.016373 | 0 |
| 10 | -0.17514 | -0.05584 | -0.04541 | 0.290522 | -0.04301 | 0 |

# Part 2 A real world regression problem – counting people

## ORIGINAL FEATURE SET ANALYSIS

In this section, the original feature is used to prediction the number of people.

Table Errors of different methods using original feature set

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | BR | LASSO | LS | RLS | RR |
| MSE | 2.618734 | **2.464461** | 3.102838 | 2.618734 | 3.118997 |
| MAE | 1.282433 | **1.256468** | 1.358444 | 1.282433 | 1.364567 |

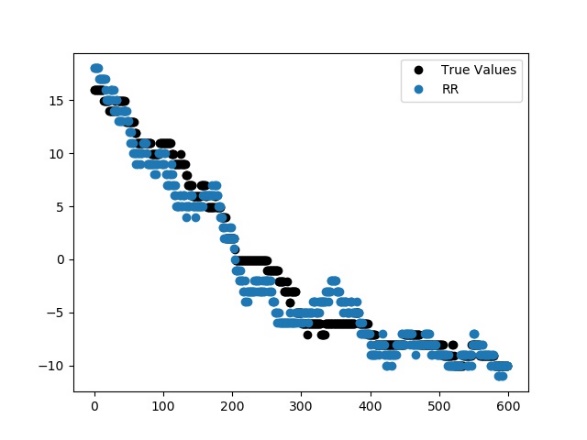
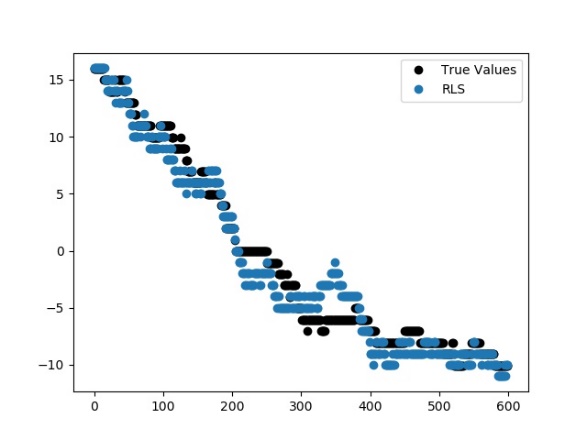
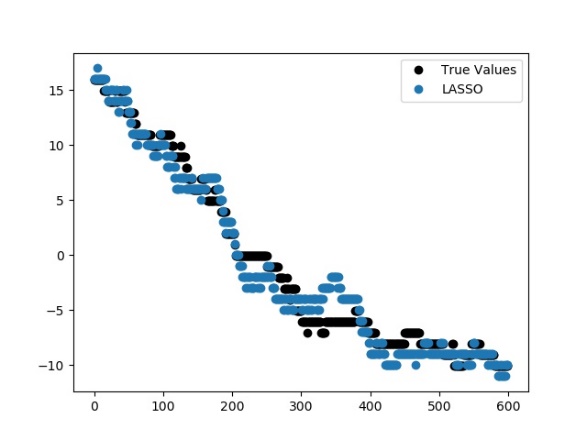
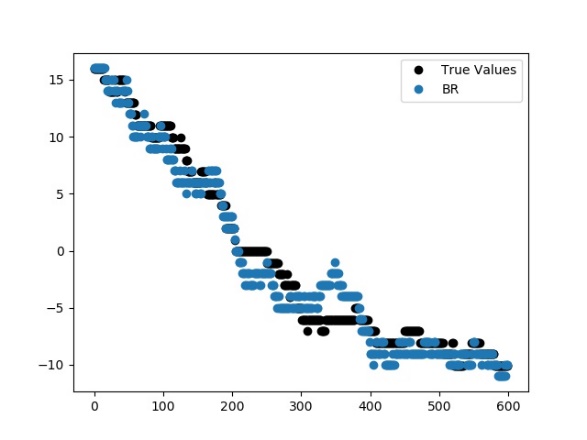
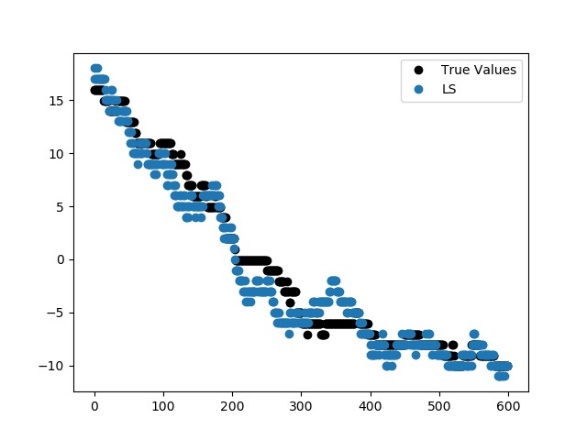
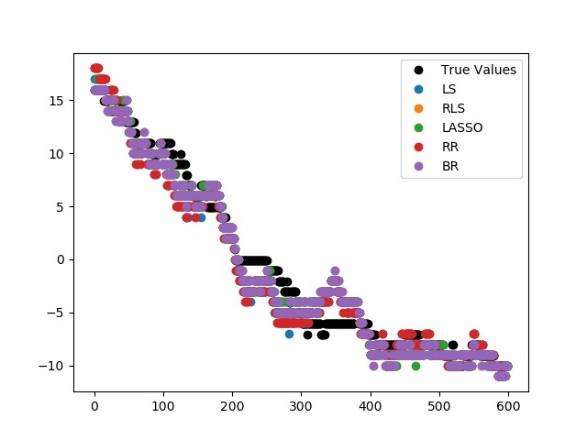


Figure Plot of predict data and true number (figure on the upper left represent different colors represent different methods)

From Table 5, we can see that LASSO contribute the smallest MSE and MAE using the original feature. From the plots, we can see that shape of prediction plots is similar across all methods, the difference between different methods is not significant. On the other hand, most wrong predictions locate in the region between -5 to 0 (sample 200-400). If we can improve the performance in this region, the performance may increase.

## Other feature transformation

In this experiment, we apply the hyperbolic tangent function (tanh) to the original features. The hyperbolic tangent function is a function shown as follows:

At the same time, we use a ploy nominal function which order is 6 to do the regression. The transformed features in LS and RR is not as good as the original ones. However, using the transformed feature set, we output a smaller MAE using LASSO while output smaller MSE using BR. The enhanced performance is obtained by increasing the order and change the transformation. Thought the result is obtained based on trial and error, we think that the tanh function as a normalizer. On the other hand, the predict result compared with the original feature set make less error in extreme values which contributes more to the MSE and MAE.

Table 6 Errors of different methods using tanh feature set

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | BR | LASSO | LS | RLS | RR |
| MSE | **2.461752** | 2.546504 | 3.679852 | 2.485101 | 3.841324 |
| MAE | 1.249837 | **1.204313** | 1.447786 | 1.238912 | 1.502059 |

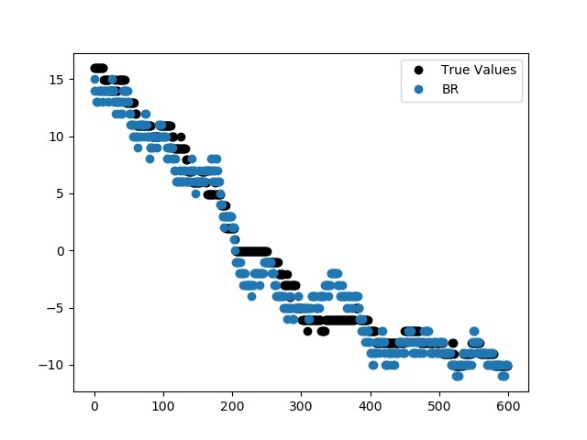
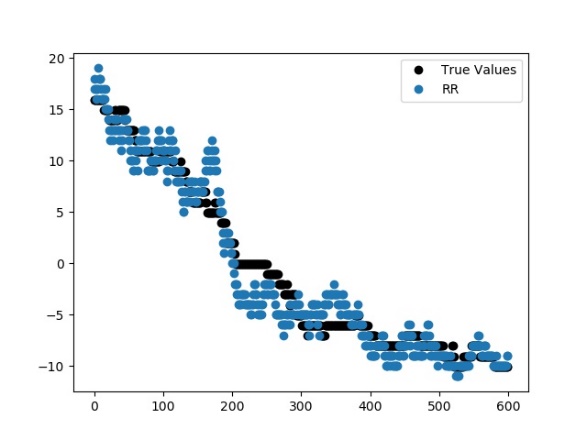
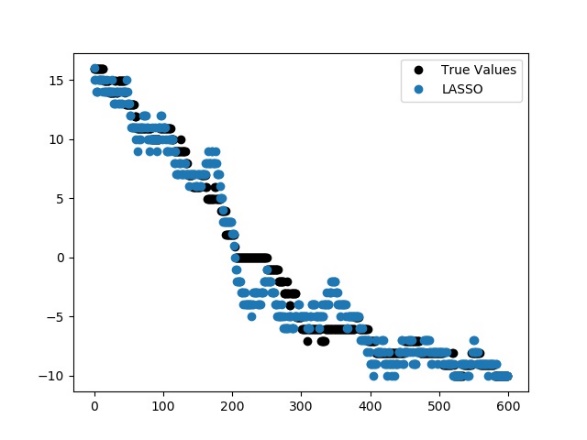
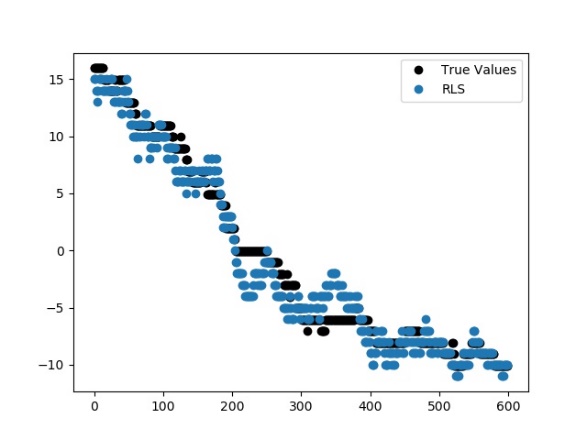
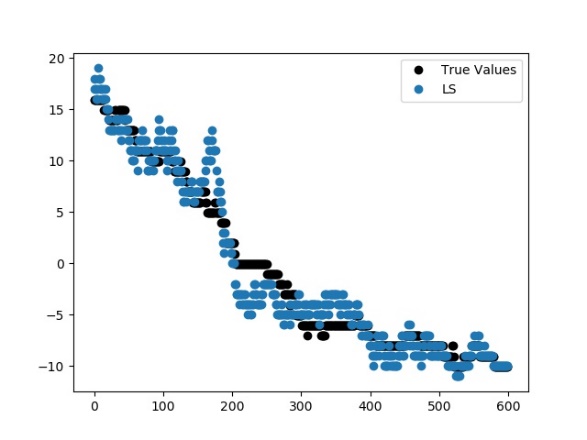
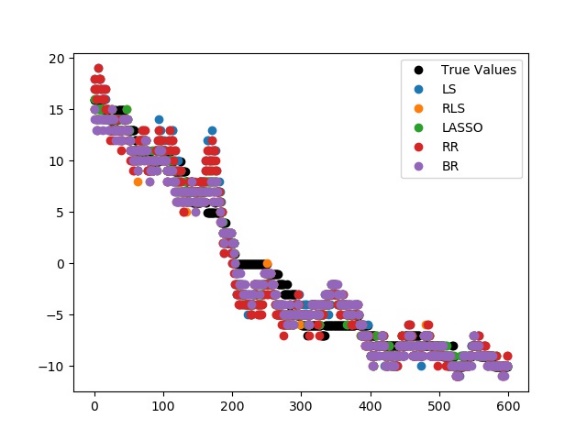


Figure Plot of predict data and true number (figure on the upper left represent different colors represent different methods)