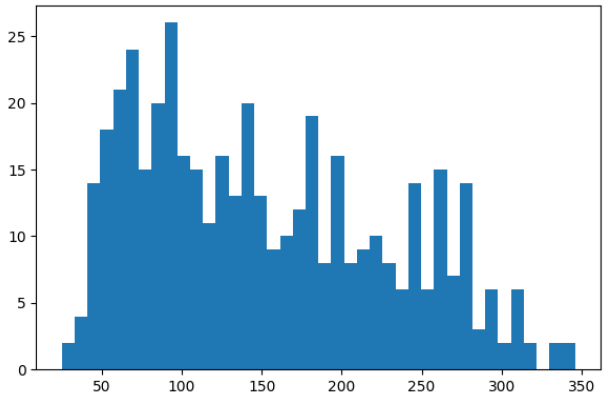
**CO544: Machine Learning and Data Mining**
**Machine Learning Lab Four**

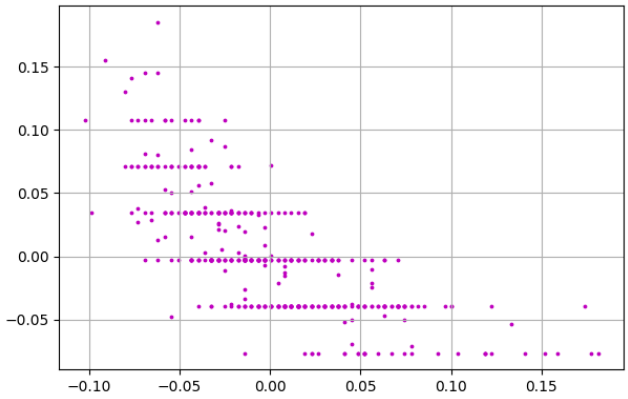
Ranage R.D.P.R. - E/19/310

## 1. Linear Least Squares Regression

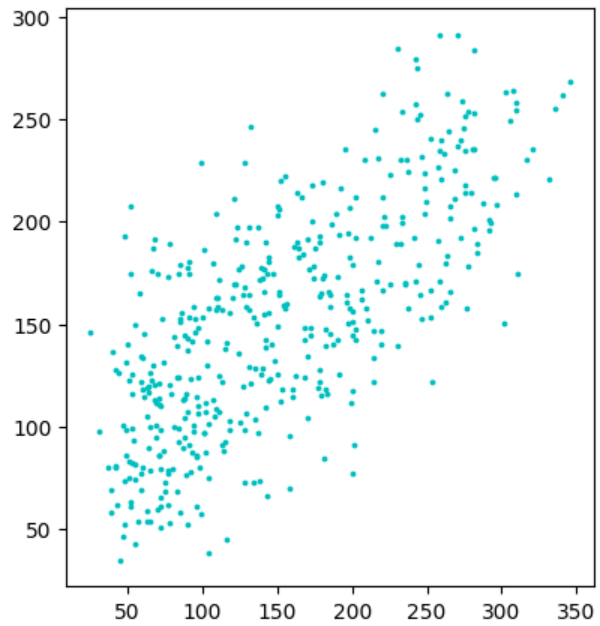
**I). Histograms of the targets and pair-wise scatters of the features in any new problem you are tasked to solve.**

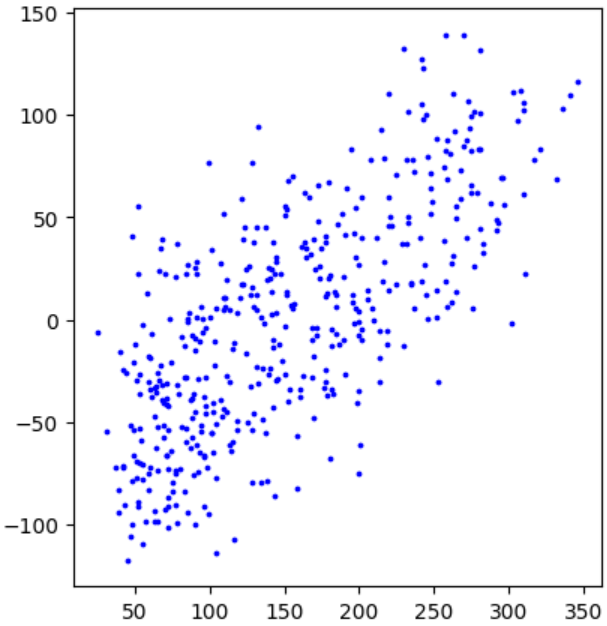
Number of samples: 442 Data shape: (442,)
Number of features: 10 Target shape: (442,)



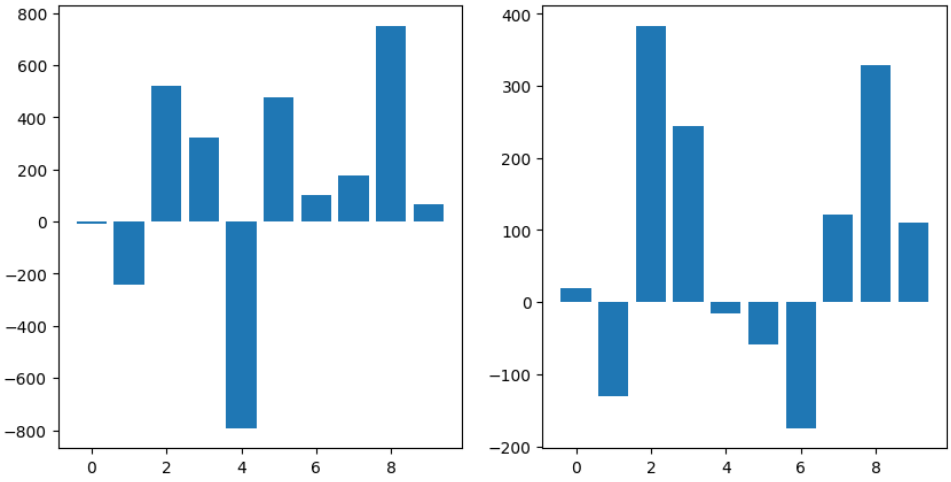

**II). Implement a linear predictor that is solved by the pseudo-inverse method**
**Solve the same problem using the linear model from sklearn and compare the results.**



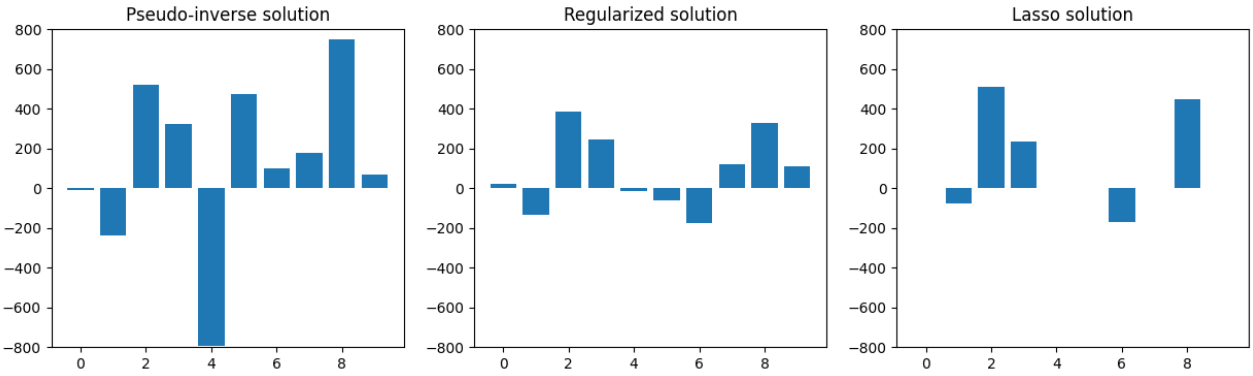
## 2. Regularization

**Derive and implement a regularized regression. Show, using two bar graphs of the weights side by side to the same scale, how the two solutions differ.**



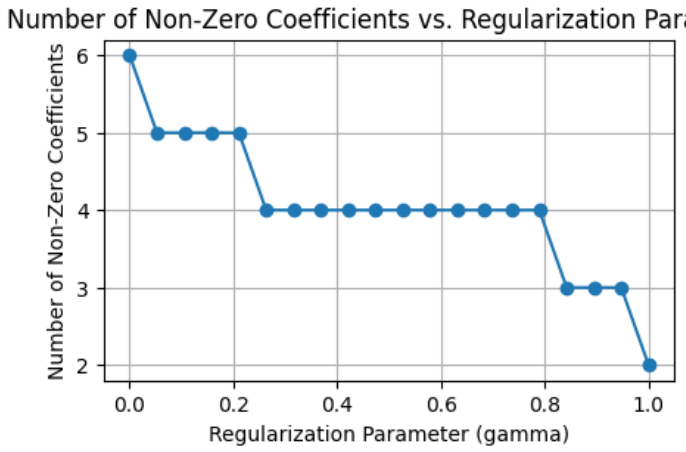

## 3. Sparse Regression

**For the Diabetes problem considered above, solve the lasso problem and plot the resulting weights as a bar graph.**




**Observe how the number of non-zero weights change with the regularization parameter γ.**

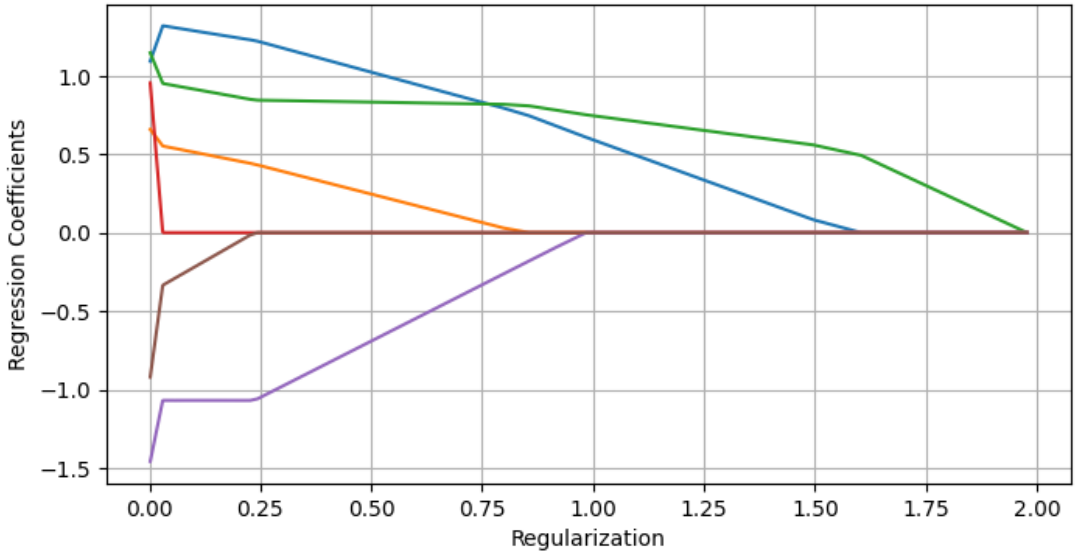
As the regularization value (γ) grows, the number of non-zero weights decrease.



**In the case of the sparse regression, would you say the features with nonzero weights are more meaningful**

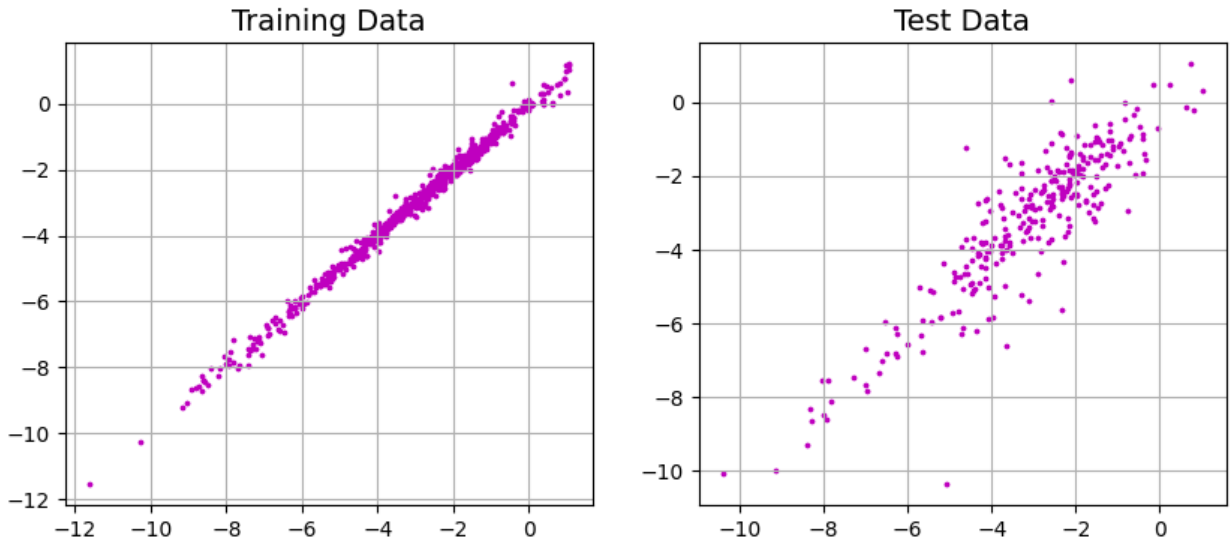
Yes, it is thought that qualities having non-zero weights are more significant or relevant. This is due to the fact that Lasso regression selects variables and applies regularization to boost the statistical model it generates interpretability and prediction accuracy.

**Regularization Path**

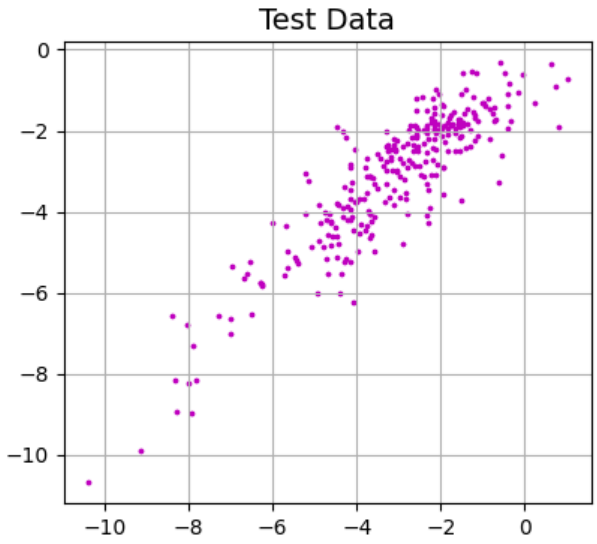



**4. Solubility Prediction**

**I). Load the data, split into training and test sets, implement a linear regression and plot the predicted solubilities against the true solubilities on the training and test sets. To facilitate comparison, draw the two scatter plots side by side to the same scale on both axes.**



**II). Implement a lasso regularized solution and plot graphs of how the prediction error (on the test data) and the corresponding number of non-zero coefficients change with increasing regularization.**


**III). If you were to select the top ten features to predict solubility, what would they be? How good is the prediction accuracy with these selected features when compared to using all the features and a quadratic regularizer?**
Top 10 features: ['SpMax4_Bh(m)', 'P_VSA_v_3', 'P_VSA_p_3', 'MLOGP', 'MLOGP2', 'ALOGP', 'ALOGP2', 'BLTF96', 'BLTD48', 'BLTA96']

MSE with all features: 1.92123251403054946
MSE with top 10 features only: 0.6781356644530957

**IV). Are you able to make any comment comparing your results to those claimed in [3] or [4]?**
The number of non-zero coefficients is higher when compared to Part 3. Additionally, both the Lasso and Tikhonov regularizations' regularization parameters have changed from Part 3 to this one.