Sensitivity of Least Squares Regressions

Names: Jaden Snell, Henry Dyer, Blake Hamilton, Iker Acha

Abstract:

Throughout this paper we plan on investigating the sensitivity of least squares regressions, particularly focusing on the robustness of results in the presence of multiple outliers. We will analyze how different types of outliers impact the coefficients and overall performance of our regression model. By means of this study, we will consider multiple pre-processing techniques to mitigate noise effects and evaluate the Least Squares Regression robustness against noise and outliers in our data sets.

By employing theoretical insights from literature alongside empirical experimentation, this study aims to provide a comprehensive understanding of noise impact on least squares regressions and proposes strategies to enhance robustness while simultaneously offering valuable insights for modeling data sets in noisy environments. Ultimately, our findings contribute to a deeper understanding of sensitivity of Least Squares Regression which might prove to be insightful for practitioners and researchers in the field.

Presentation of Introductory Information:

We plan on introducing the concepts of noisy, linear least squares, non-uniform noise and heteroscedasticity, and noisy regression through simple examples.

For noisy, linear least squares, we will write an algorithm to calculate least squares regression, then find noisy data and calculate least squares regression for $y=x^2$ on an interval $[a,b]$. We will calculate the regression in the forms $f_1=ax^2$, $f_2=ax^2+bx+c$, $f_3=ax^4+bx^3+cx^2+dx+e$ and find the approximation that's most accurate in all combinations of x and y. Our analysis will include when the noise is the worst and we'll repeat with the interval $[b,a+b]$. After running the regression with noise, we will remove noise and run the regression while comparing different intervals and analyzing the performance.

For the case of non-uniform noise and heteroscedasticity, we will derive an estimator that gives the optimal solution to the weighted least squares problem. Next we will build a heteroscedastic data set and determine regression for the dataset using standard least squares and weighted least squares regression. We will then analyze those results and find the most accurate approximations.

For noiseless regression, we will generate a noiseless set of points along $y=x^2$ and add an outlier to the data set. Then we will calculate the line of best fit through different forms of regression (f_1,f_2,f_3). We will rerun the experiment with an increased and decreased number of data points and analyze the effects. Finally we will analyze the effect of outliers and determine when they are necessary to remove and how to deal with outliers to improve the models. This analysis will consist of using other norms to measure the error vector.

Independent Extension:

For our independent extension we will first explore how total least squares effects the sensitivity of noise within the regression. Secondly, we will explore how correlated noise between the different measurements affects the overall performance of least squares. This will

be done by sampling from a multivariate distribution for dimension *n* for *n* discrete points instead of sampling from *n* independent univariate so we get a correlation between the residuals. Finally, if we have time we may explore heteroskedasticity among the residuals as well as errors among the *x* values, not simply the *y* values.

Timeline:

3/13-4/6:

All members complete the project proposal by April 3. The rest of the week will consist of meeting with Alex or Eduardo and further planning the presentation of the introductory information.

4/7-4/13:

Complete the introductory material background and mathematical proofs for the final report focusing on noisy linear models. Create thorough models and graphs in Python to display the noisy linear theorems that we have discussed in class.

4/14-4/20:

Focus on non-uniform noise and heteroscedastic models for further introductory background material. As above, create models and graphs in Python to represent the work done and give the reader clear background.

4/21-4/27:

Work on the total least squares extension. As above, create models and give mathematic explanations and proofs sufficient to show the reader our progress. Start Regularized least squares.

4/28-5/2:

Complete Regularized least squares and formalize all findings in the project. Polish all code, graphs, and mathematical work to be placed in the final document.

5/3-5/7:

Finalize report and prepare for presentation.

Work Distribution

Blake:

- Lead on Noisy Linear Models: Develop introductory material and mathematical proofs, focusing on noisy linear models
- Lead on Total Least Squares Extension: Take charge of developing models, providing explanations, and proofs for total least squares extension.

Iker:

- Project Coordination: Ensure timely completion of project proposal and facilitate meetings for planning and progress updates.
- Documentation and Support: Assist in documenting progress, provide support to team members, and finalize the report.

Jaden:

- Focus on Heteroscedastic Models: Collaborate on introductory material, especially on non-uniform noise and heteroscedastic models.
- Lead on Graphical Representation: Create models and graphs in Python to illustrate theoretical concepts.

Henry:

- Collaboration on Introductory Material: Work with Blake on developing background material and mathematical proofs.
- Contributor to Total Least Squares Extension: Collaborate with the team on total least squares extension and finalization of findings.