Project 2: Sensitivity of Least Squares Regressions

University of Colorado Boulder
Department of Applied Mathematics

1 Project Summary

In many applications the data is noisy; i.e. the data is not exactly the data you want to fit. In this project, your task is to explore the sensitivity of the discrete least squares approximation to the noise in the data. Does the accuracy depend on the number of data points and/or the size of the noise? Are there other things that come into play? Find theory in the literature to support your findings. Possible options for the independent part include techniques for pre-processing data to make it less noisy and least squares techniques that make it more robust to noise and outliers (e.g. smoothing).

2 Project Background

2.1 The Method of Least Squares Regression

In any experimental study, a common goal is to develop or test a model for some observable phenomena based on recorded observations. To restate this goal mathematically, we often aim to build a continuous form, f(x), that accurately describes a discrete and finite set of data points, $\{(x_1, y_1).(x_2, y_2), ...(x_m, y_m)\}$. One method for building these models is the method of least squares regression, which aims to find an approximation f(x) as a linear combination of other functions $\{g_i(x)\}_{i=1}^n$ or

$$f(x) = c_1g_1(x) + ... + c_ng_n(x).$$

For example you might want to approximate f(x) as a second order polynomial. To do this you would select $\{g_i(x)\}_{i=1}^n = \{1, x, x^2\}$, and you would write f(x) as

$$f(x) = c_1 + c_2x + c_3x^2$$

and then solve for the coefficients c_1, c_2, and c_3.
This is done by minimizing the total squared error between the continuous model (f(x)) and the discrete data set. This is achieved by solving the equation

$$||\boldsymbol{Gc} - \boldsymbol{y}||_2^2 = 0$$

where $\boldsymbol{y}$ is a vector of the independent variable measurements y_i, $\boldsymbol{c}$ is a vector of the function weightings c_i and $\boldsymbol{G}$ is of the form

$$\boldsymbol{G} = \begin{bmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_m) & \dots & f_n(x_m) \end{bmatrix}.$$

Since we often do take exactly as many more measurements (the rows of $\boldsymbol{G}$) than we have variables in our model to (the columns of $\boldsymbol{G}$), $\boldsymbol{G}$ is often a non-square matrix. This means that the vector $\boldsymbol{c}$ that provides the least squares solution often does not exist, and must instead be approximated using a matrix pseudo-inverse ($(\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T$). In cases where G is ill-conditioned ($\kappa(\boldsymbol{A})$ is large) this operation can result in approximations of $\boldsymbol{c}$ that att

2.2 Getting Noisy

In real life most experimental data is noisy and as a result contain false information. Reword the previous sentence so it is more expository. For example, most commercially available resistors are manufactured with a tolerance of $\pm 1 - 5\%$. This means that in many applications, the data set that we are given to build a model do not only include information from the physics, system, etc. that we are trying to model. This is the reason why it is important for us to understand both how noise effects the results of the methods we are using to build our model. With this information, we can do our best to minimize the effects of noise without compromising the integrity of the data and/or make an educated decision about what is the best way to build a model given a set of data. You will be begin by investigating the effect of noise in the data when using a least squares

2.2.1 Questions to Investigate

1. Write an algorithm to calculate the least squares regression of the form f(x) for the data set $\{x_i, y_i\}_{i=1}^m$.

2. Using the algorithm you made above, calculate the least squares regression for some noisy data for the a quadratic function $y = x^2$ on the interval $[a, b]$. Calculate regressions of the form $f_1(x) = ax^2$, $f_2(x) = ax^2 + bx + c$, and $f_3(x) = ax^4 + bx^3 + cx^2 + dx + e$. Which approximation is the most accurate to the noisy data set? Which approximation is most accurate to the actual model? You should run this experiment where the noise is present in only the x-data, only the y-data, and in both. When does the noise create the biggest issue?

3. Compare each of the model you created above to $f(x) = x^2$ on the interval [b, b+a]. Which model performs the best?

4. Now remove the noise from your data set and run your least squares regression for f_1, f_2 and f_3 and compare their performance over the intervals [a,b] and [b, b+a]. Which method performs the best? If anything changed what happened? What does this mean about the performance of the least squares regression when noise is present? What do we need to know about our system in order to get an accurate model?

Maybe add an exercise where they use the polynomial to fit a more complicated function with noise. Also, you do not specify how to make the noise. There are different ways. They can explore that as well.

2.3 Non-Uniform Noise and Heteroscedasticity

Often a sensor is used to measure different experimental values. This is the *data* that is used in the modeling process. Most sensors have operating ranges were they have a consistent error in their measurements on range of values, and a different, sometimes variable error, for values outside of that range. Either reword or provide an example for the last sentence. What this translates to is some data measurements are more accurate than others. Figure **??** illustrates an example of a *heteroscedastic* data set. where what happens? Why are they looking for in the figure? You never define heteroscedastic.

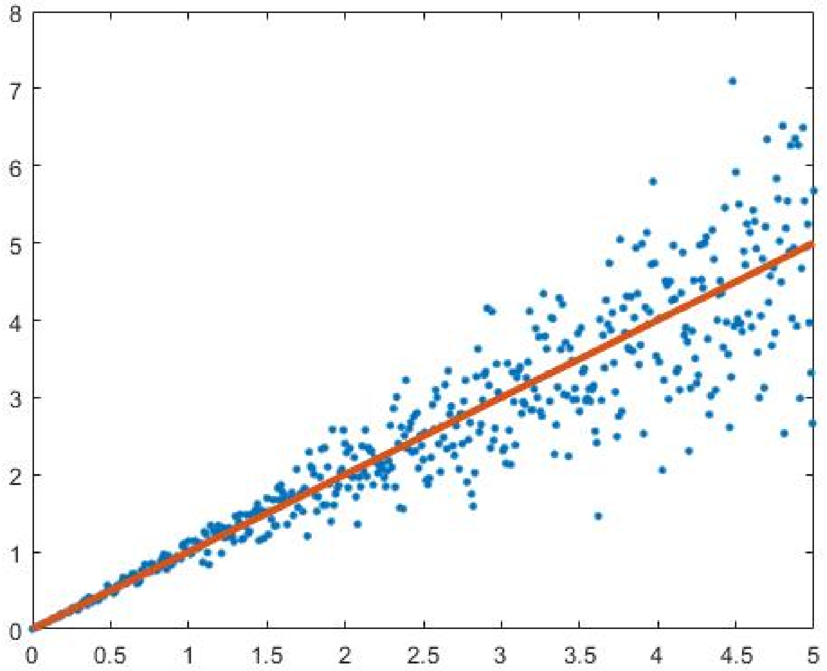

Figure 1: Illustration of a data set generated to approximate the line $y = x$. This illustrates the phenomena of heteroscedasticity.

What this (what is this?) results in is a collection of data where some of the data is more closely correlated with the actual physics of a given problem than other data points. Since the more correlated points have less error (in what sense?) they effect the least squares loss function less. There are several ways to deal with this issue. The first option is to discard the data with high error, keeping only the data points with low error. (How do they know the error?) While this will obviously "fix" the (what is the name of the issue?) issue, it is not often practiced since even noisy data is still representative of a system being measured and data collection is expensive in terms of time and/or money. Furthermore, if you are approximating something that is more complicated than a straight line it is possible to exclude key details from the model by throwing away data.

Since excluding data is a crude method that has many undesirable properties, a superior method is needed. An alternative is to use a weighted least squares technique for fitting the data. With the weighted least squares method, each residual is given a weight proportional to some weighting factor w that correlates with the magnitude of the noise. Once an appropriate weighting factor has been chosen the new loss function for the least squares regression is given as

$$\underset{\boldsymbol{x}}{\operatorname{argmin}} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_w^2,$$

where $\| \cdot \|_w$ is defined as follows:

$$||\boldsymbol{x}||_w = \sqrt{c_1 x_1^2 + c_2 x_2^2 + \ldots + c_n x_n^2}. \tag{1}$$

(Is the weight missing above ?)

2.3.1 Questions to Investigate

1. Recall that any vector norm can be written as $\boldsymbol{u}^T\boldsymbol{C}\boldsymbol{u}$. (What are $\boldsymbol{u}$ and $\boldsymbol{C}$?) Using the definition of the vector norm and equation (**??**), derive the estimator that gives the optimal solution to the weighted least squares problem.

2. Build a heteroscedastic data set. Determine a regression for the data set using standard least squares and weighted least squares. Which regression most accurately approximates the actual model underlying the data? Students probably need more structure in this question.

2.4 Being out there

Another key issue that arises when collecting data is the possibility of a poor measurement. For example, if a table got kicked in a biology lab disturbing the items being measured. Mathematicians refer to this ”poor measurement” as an outlier. In an introduction to statistics class, students are sometimes told that an outlier can be problematic in their data set because the large value of an outlier can throw off the mean of the data leading to a misunderstanding of the story that a data set is telling. What about the case of a least squares regression? What are the effects?

2.5 Questions to investigate

1. Generate a noiseless set of points that fall along the line $y = x^2$ and add an outlier point to the data set. Calculate the line of best fit using the models $f_1(x) = ax^2$, $f_2(x) = ax^2 + bx + c$, and $f_3(x) = ax^4 + bx^3 + cx^2 + dx + e$. What do you observe in the regression lines? Which forms are most robust against the outlier?

2. Increase and decrease the number of data points you used in the previous question, does this effect the accuracy of the approximation?

3. Why do outliers have the effect they do on the least-squares regression? Why does increasing and decreasing the number of data points being fit effect the regression like it does?

4. One practice with outliers is to remove them from the data set before analyzing it, how bad does an outlier need to be to throw off a data set?

5. What options do we have when it comes to outliers to improve our models? What are the effects of using other norms to measure the error vector?

3 Software Expectations

You are expected to implement all of the algorithms for this project yourself. Please see the instructor for guidance if your independent direction may require the use of packaged software.

4 Independent Directions

Possible next steps for this project include, but are not limited to:

1. What reprocessing techniques can we apply to our signal to reduce noise? Are some techniques more effective than others?

2. How can we reformulate the Least Squares Technique to increase robustness against noise and outliers?

3. How can we reformulate the Least Squares Technique to encourage sparsity in regressive models?

4. How does noise effect the nonlinear formulation of least squares?

5 Helpful Sources

1. Golub, Van Loan, Matrix Computations, Chapters 5 (Ordinary Least Squares) and 6 (Modified Least Squares)

2. Demmel, Applied Numerical Linear Algebra, Chapter 3: Linear Least Squares problems

3. Hastie, Tibishirani, The Elements of Statistical Learning, Chapter 3: Linear Methods for Regression