A web-based application for the visual exploration of regression outliers, high leverage points, and influential points

Christopher J. Casement
casementc@gmail.com
Department of Mathematics
Fairfield University
Fairfield, Connecticut, 06824
USA

Abstract

Regression outliers, high leverage points, and influential observations are commonly taught in statistics courses in which regression methods are covered – from the introductory to the advanced levels. When covering these topics, it is typical to discuss various methods that exist for determining whether a point is one of these types – such as Cook's distance or DFFITS. Yet while these topics are routinely covered, it can be difficult for instructors to graphically illustrate whether observations are outliers, high leverage points, or influential observations based on the methods and their associated criteria. To address this shortcoming, in this article, a free, web-based app created by the author, which focuses on influential and related observations, is described, with intended usage by both faculty and students. Faculty feedback, which highlights the accessibility, ease of use, and usefulness of the app, is also discussed.

1. Introduction

Regression outliers, high leverage points, and influential observations are topics commonly taught in statistics courses that cover regression. A point is deemed a regression outlier if it falls outside of the overall trend or pattern of the data (particularly in the y-direction), and a high leverage point if it is extreme in the x-direction (typically relative to the mean of the x-values). A point is then considered to be potentially influential if it is both a regression outlier and a high leverage point. Various measures and cutoffs are used to determine if a value is one of the aforementioned types. For instance, the standard deviation of the residuals and studentized residuals are commonly used to assess whether a point is a regression outlier; the leverage statistic (also known as the hat value) is used for high leverage points; and Cook's distance, DFFITS, and DFBETAS can be used to examine whether a point is potentially influential.

While teaching regression outliers, high leverage points, and influential points is common, it is perhaps most beneficial to pair them with a graphical tool when explaining them in a simple linear regression setting. This is due to the fact that it has been found that visualizations can boost the ability of students to understand new concepts [2, 4, 11]. To facilitate these efforts, the author of the article created a free, web-based application that implements the aforementioned methods for detecting potential points of the three types.

The article proceeds as follows. In Section 2, the use of real data in the classroom is discussed, as are recommended characteristics for evaluating statistical tools. Then, in Section 3, a description and example of the app is provided. Next, in Section 4, a survey given to faculty who teach statistics regularly is described and the results discussed. Lastly, the article is summarized in Section 5.

2. Background

Numerous articles have focused on the importance of using real data, technology, and active learning in the classroom. For instance, Garfield and Ben-Zvi [6], Garfield and Everson [7], and Neumann, Hood, and Neumann [10] focus on statistical reasoning in introductory statistics courses, while others, including Rumsey [15] and Singer and Willett [18] have discovered that students participate more when real world situations and data are utilized.

Of course, when real world data is involved, oftentimes it is not as ‘clean’ as one might hope for. For instance, the violation of assumptions for a particular statistical method (e.g., due to the presence of outliers) can lead to the potential for misleading or even invalid conclusions to be drawn without proper care being taken. As far as regression is concerned, regression outliers, high leverage points, and/or influential points can lead to such issues. To this end, anyone running a regression analysis should fully understand these types of data points and how they can produce such issues.

A plethora of tools exist that enable users to run regression analyses, including point-and-click programs and programming languages. Point-and-click programs include JMP, SPSS, JASP [8], StatCrunch [12], and Rguroo [17], among others. While some of these programs enable the calculation of certain statistics for assessing whether a point might be a regression outlier, high leverage point, or influential point, the programs typically calculate the measure of interest and report its values, such as in the form of a new column in the original dataset, rather than displaying them graphically. Or, if the program does plot the values, it might do so in a separate plot. Some programming languages also enable users to run regression analyses, such as R, Python, MATLAB, and Stata (while not a programming language itself, it has programming languages built in), among others. While the calculation of the measures of interest is not always particularly burdensome using these languages, a specialized knowledge of the languages is required in order to accomplish various things that the proposed app has built in automatically, such as:

- interactively moving points in a scatterplot (which is not even doable using some languages) and viewing the potential impact on regression statistics,
- adding statistic values (e.g., Cook’s distance or DFFITS values) to a scatterplot, and
- quickly and easily identifying which observations in a dataset are potential outliers, high leverage points, or influential points based on the statistic of interest.

In fact, none of the aforementioned tools – point-and-click programs or programming languages – (to the author’s knowledge) provides a ‘ready-to-go’ tool for the visual exploration of potential regression outliers, high leverage points, or influential points via the commonly-used statistics discussed in Section 1, all in a single scatterplot. Of course, one could create an interactive tool (like the author did) using certain programming languages such as R, but that would require specialized knowledge of the language, which is not expected of the vast majority of students learning regression or instructors teaching it.

In order to evaluate technological tools, Biehler [1], McNamara [9], and Repenning [14] provide attributes to consider. In particular, McNamara focuses on ten key characteristics for evaluating tools for statistical computing. However, due to the proposed app’s focus on the plotting and fitting of data, the six characteristics included in Table 1 – all of which are possessed by the app – are of particular interest. The table also explains how the app meets each characteristic.

Table 1. The proposed application in relation to characteristics recommended by McNamara [9].

Characteristic	How the App Possesses the Characteristic
“Be accessible”	The app is free to use and simply requires a web browser.
“Provide easy entry”	The app has a point-and-click interface and does not require any programming background.
“Privilege data as a first-order object”	The app requires users to input data, with the resulting scatterplot and regression output focused on that data.
“Support exploratory and confirmatory analysis”	The app enables users to explore their data via the scatterplot and associated regression statistics. It also provides users with desired statistics for determining whether a point might be an outlier, high leverage point, or influential point.
“Allow for flexible plot creation”	The app automatically creates a scatterplot for users and allows users to customize the axes.
“Be interactive”	The app utilizes the *plotly* [16] package in R [13], enabling users to manually move points displayed in the scatterplot.

3. Description and example of the application

A description of the *Influential Points* application is now provided. The app was created using the *shiny* package [5] in R [13] and is free to access at https://educationapps.shinyapps.io/InfluentialPoints. The app utilizes various R packages, but most notably it uses *rio* [3] for flexible data importing and *plotly* [16] for interactive plotting. Strengths of the app are now discussed, followed by an example of its usage.

3.1 Strengths of the app

The app possesses various strengths, including the following:

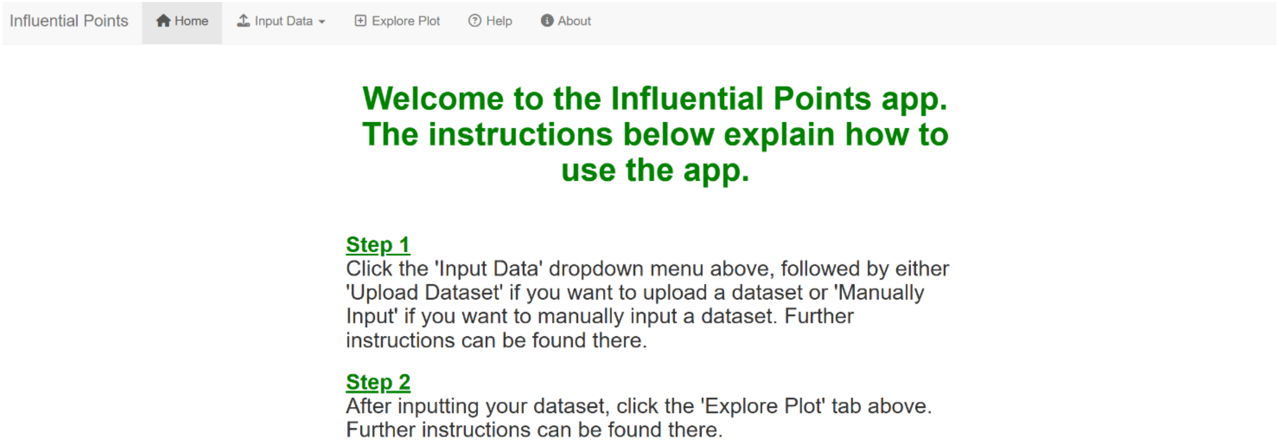
- The app is accessible via the internet. In fact, it can be accessed via any device with internet access (e.g., laptops, tablets, and even smartphones).
- The app is free to use by both faculty and students.
- Users have the option to either manually input data or upload a dataset using files of various common types – e.g., .csv, .xlsx, .sav, and .txt, among others.
- The app displays values for common methods of detecting regression outliers, high leverage points, and influential points, including the standard deviation of the residuals, studentized residuals, leverage values (i.e., hat values), Cook’s distance, DFFITS, and DFBETAS.
- Users can manually move individual points on the scatterplot to examine the potential impact of outliers, high leverage points, or influential points on regression output. For instance, they can see how the slope, y-intercept, correlation coefficient, r-squared, standard deviation of

the residuals, and the p-value (when testing the slope) do or do not change as a result of one of these types of points.

3.2 Example using the app

An example of the app's usage is now provided using real-world data collected in an introductory statistics course. The dataset contains data on houses, including the selling price, house size, number of bedrooms, and number of bathrooms, among other variables. The focus of the example is on regressing selling price (measured in U.S. dollars) on house size (measured in square feet).

When the user opens the app, they are immediately taken to the 'Home' tab, which displays instructions for how to use the app, as can be seen in Figure 1.



Welcome to the Influential Points app. The instructions below explain how to use the app.

Step 1
Click the 'Input Data' dropdown menu above, followed by either 'Upload Dataset' if you want to upload a dataset or 'Manually Input' if you want to manually input a dataset. Further instructions can be found there.

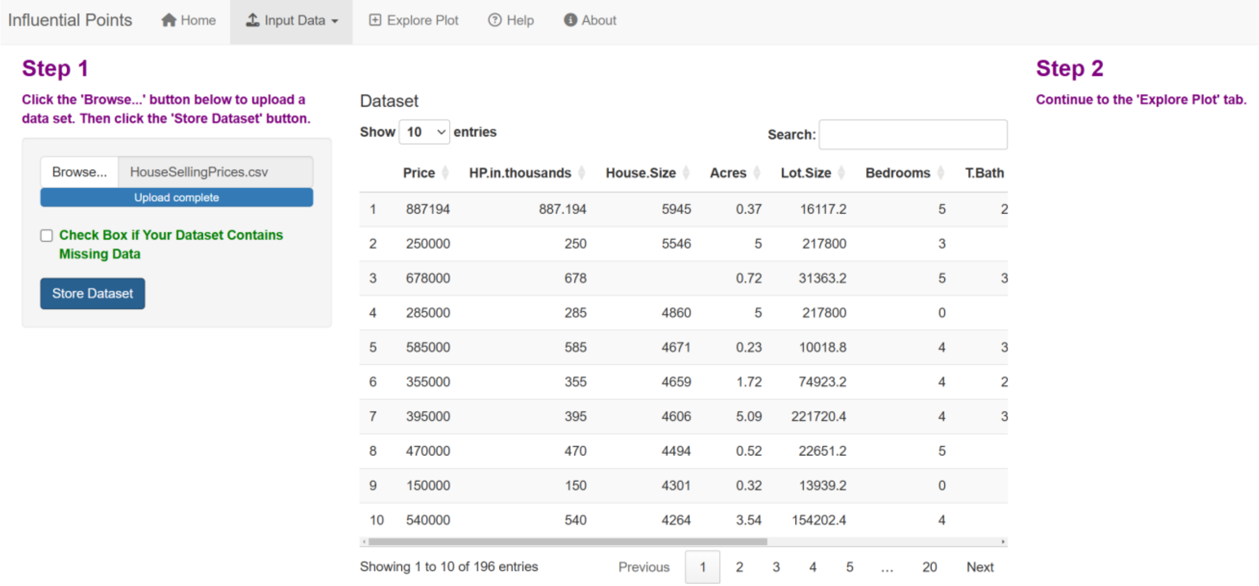
Step 2
After inputting your dataset, click the 'Explore Plot' tab above. Further instructions can be found there.

Figure 1. After opening the app, the user starts in the 'Home' tab, which contains instructions for how to navigate the app.

After reading the instructions, the user should click on the 'Input Data' dropdown menu at the top of the app, where they can either upload a dataset or manually input data. In this case, the user can quickly and easily upload their existing dataset. As Figure 2 displays, the user has clicked the "Browse…" button to select the dataset from their computer and then clicked the blue 'Store Dataset' button to store the dataset in the app for further usage.

Once they have uploaded their dataset, the user then navigates to the 'Explore Data' tab found at the top of the app. There, they start by selecting the two variables of interest (X and Y) from the respective dropdown menus, and they then click the blue 'Make Plot' button to make a scatterplot of the selected data. Figure 3 displays the resulting scatterplot when regressing selling price on house size, as well as the regression output, which includes the y-intercept and slope of the regression line, the correlation coefficient, r-squared, the standard deviation of the residuals, and the p-value for testing the slope of the regression line. Additionally, users have the option of customizing the x-axis and y-axis labels, as is done in Figure 3.

When examining the scatterplot, the user should look for potential regression outliers, high leverage points, and influential points, as those types of observations are the main focus of the app.



Step 1

Click the 'Browse...' button below to upload a data set. Then click the 'Store Dataset' button.

□ Check Box if Your Dataset Contains Missing Data

Step 2

Continue to the 'Explore Plot' tab.

Dataset

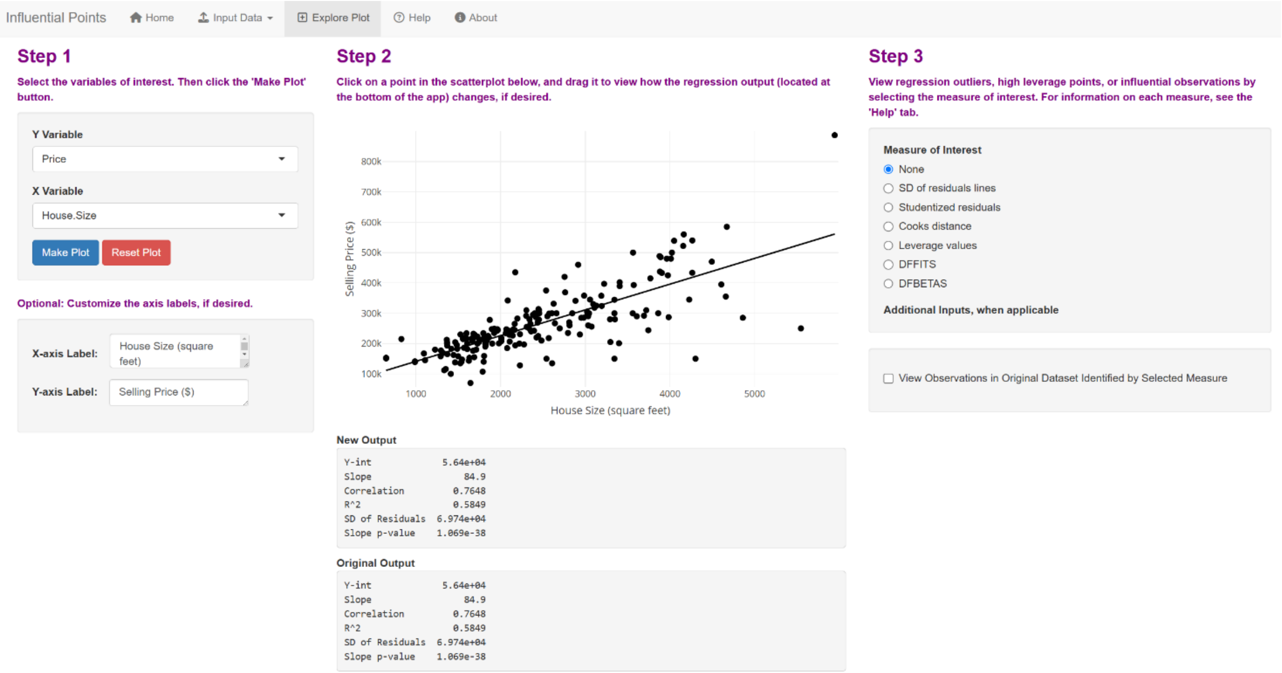
Show 10 entries

Search:

	Price	HP.in.thousands	House.Size	Acres	Lot.Size	Bedrooms	T.Bath
1	887194	887.194	5945	0.37	16117.2	5	2
2	250000	250	5546	5	217800	3	
3	678000	678		0.72	31363.2	5	3
4	285000	285	4860	5	217800	0	
5	585000	585	4671	0.23	10018.8	4	3
6	355000	355	4659	1.72	74923.2	4	2
7	395000	395	4606	5.09	221720.4	4	3
8	470000	470	4494	0.52	22651.2	5	
9	150000	150	4301	0.32	13939.2	0	
10	540000	540	4264	3.54	154202.4	4	

Showing 1 to 10 of 196 entries

Figure 2. The user can upload a dataset in the 'Input Data' tab. After the user clicks the 'Store Dataset' button, the uploaded dataset is displayed in the app.



Step 1

Select the variables of interest. Then click the 'Make Plot' button.

Y Variable: Price

X Variable: House.Size

Optional: Customize the axis labels, if desired.

X-axis Label: House Size (square feet)

Y-axis Label: Selling Price ($)

Step 2

Click on a point in the scatterplot below, and drag it to view how the regression output (located at the bottom of the app) changes, if desired.

New Output

```
Y-int           5.64e+04
Slope               84.9
Correlation       0.7648
R^2               0.5849
SD of Residuals 6.974e+04
Slope p-value   1.069e-38
```

Original Output

```
Y-int           5.64e+04
Slope               84.9
Correlation       0.7648
R^2               0.5849
SD of Residuals 6.974e+04
Slope p-value   1.069e-38
```

Step 3

View regression outliers, high leverage points, or influential observations by selecting the measure of interest. For information on each measure, see the 'Help' tab.

Measure of Interest

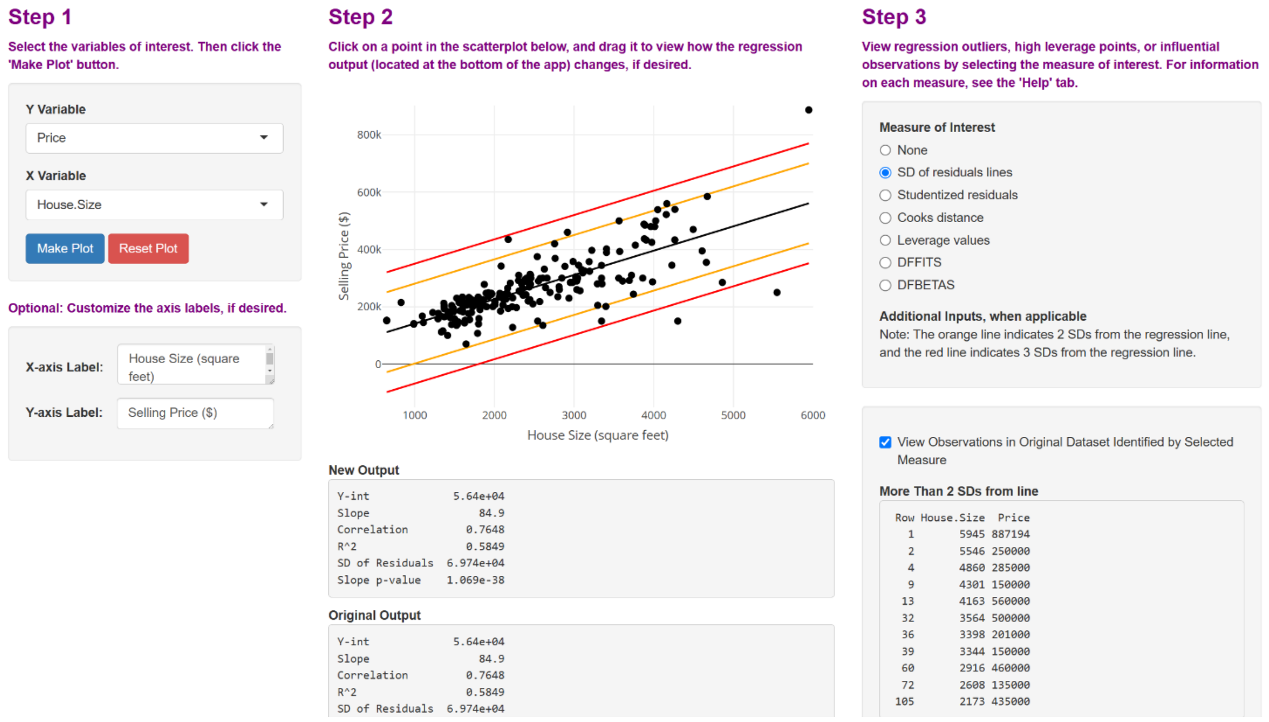
- ● None
- ○ SD of residuals lines
- ○ Studentized residuals
- ○ Cooks distance
- ○ Leverage values
- ○ DFFITS
- ○ DFBETAS

Additional Inputs, when applicable

□ View Observations in Original Dataset Identified by Selected Measure

Figure 3. A scatterplot of the user's data can be found in the 'Explore Plot' tab. Below the plot, regression statistics such as the y-intercept and slope of the regression line, correlation coefficient, r-squared, standard deviation of the residuals, and p-value (when testing the slope of the line) are printed. Additionally, the user has the option of customizing the axis labels, as is done here.

However, while a visual assessment can be helpful, determining whether an observation is a regression outlier, high leverage point, or influential point in this way is a subjective approach. For a more objective method, people analyzing data oftentimes turn to statistical measures. For instance, when assessing whether an observation is a regression outlier, two commonly-used statistics are the standard deviation of the residuals and studentized residuals, both of which are built into the app. To visually assess observations using the standard deviation of the residuals, the user should click the button corresponding to 'SD of residuals lines' for the measure of interest found in 'Step 3.' After the user has made this selection, lines are plotted in the scatterplot at values that are two (orange) and three (red) standard deviations (of the residuals) from the regression line. Any observation that falls beyond the preferred set of lines (i.e., above the orange/red lines located above the regression line or below the orange/red lines found below the regression line) is a potential regression outlier. Of course, users likely want to be able to easily determine *which* observations in the dataset are potential outliers (i.e., the row numbers). In fact, the app provides users with this option. The user can simply check the checkbox labeled 'View Observations in Original Dataset Identified by Selected Measure,' which is found in the bottom-right-hand corner of the app. Figure 4 displays the scatterplot of the housing data with the orange and red lines added, as well as the row numbers of the observations that are deemed potential outliers by falling more than two standard deviations from the regression line. While not shown in Figure 4, the app also prints the observations that are more than three standard deviations from the regression line in the bottom-right-hand corner.



Step 1

Select the variables of interest. Then click the 'Make Plot' button.

Y Variable: Price

X Variable: House.Size

Optional: Customize the axis labels, if desired.

X-axis Label: House Size (square feet)

Y-axis Label: Selling Price ($)

Step 2

Click on a point in the scatterplot below, and drag it to view how the regression output (located at the bottom of the app) changes, if desired.

New Output

```
Y-int           5.64e+04
Slope               84.9
Correlation       0.7648
R^2               0.5849
SD of Residuals 6.974e+04
Slope p-value   1.069e-38
```

Original Output

```
Y-int           5.64e+04
Slope               84.9
Correlation       0.7648
R^2               0.5849
SD of Residuals 6.974e+04
```

Step 3

View regression outliers, high leverage points, or influential observations by selecting the measure of interest. For information on each measure, see the 'Help' tab.

Measure of Interest

○ None
● SD of residuals lines
○ Studentized residuals
○ Cooks distance
○ Leverage values
○ DFFITS
○ DFBETAS

Additional Inputs, when applicable

Note: The orange line indicates 2 SDs from the regression line, and the red line indicates 3 SDs from the regression line.

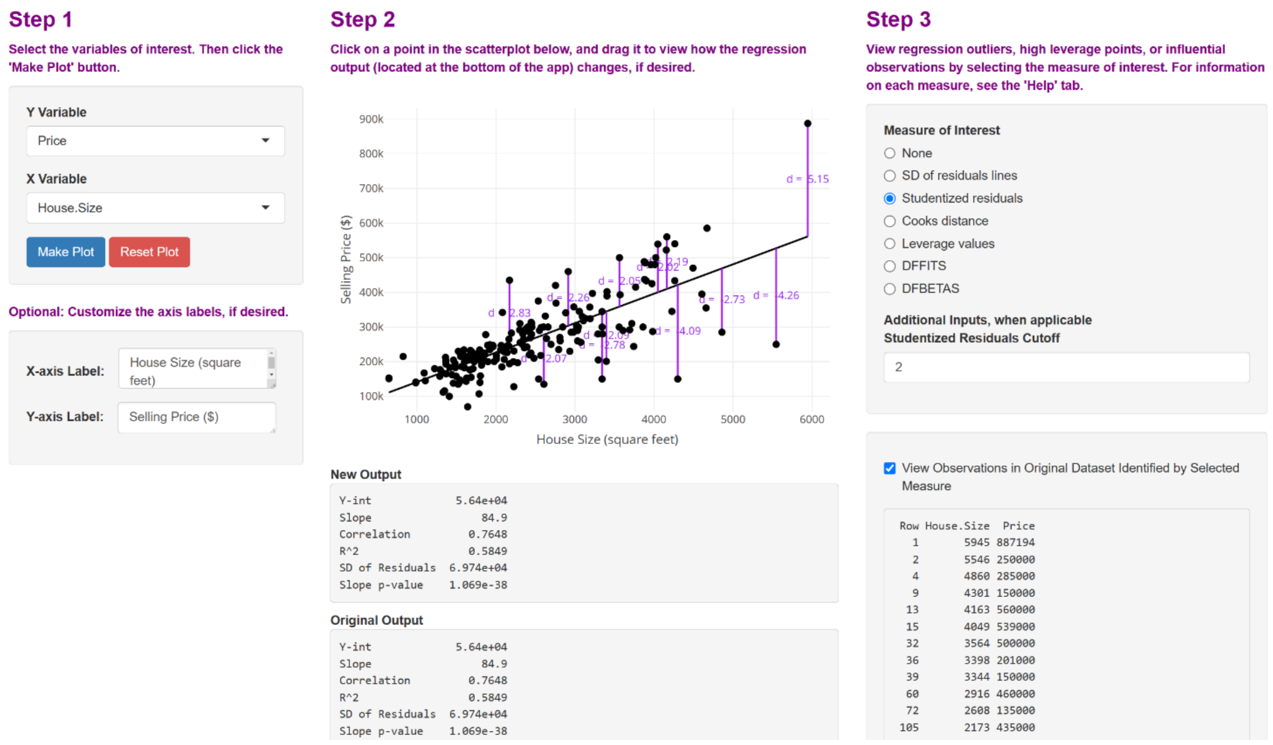
☑ View Observations in Original Dataset Identified by Selected Measure

More Than 2 SDs from line

```
Row House.Size  Price
  1       5945 887194
  2       5546 250000
  4       4860 285000
  9       4301 150000
 13       4163 560000
 32       3564 500000
 36       3398 201000
 39       3344 150000
 60       2916 460000
 72       2608 135000
105       2173 435000
```

Figure 4. When the user clicks the 'SD of residuals lines' button in 'Step 3,' lines are superimposed at two (orange) and three (red) standard deviations (of the residuals) from the regression line. Here, 11 observations are potential regression outliers, as they lie more than two standard deviations from the regression line. Users can view which observations those are in the bottom-right-hand corner.

The other common statistic used for regression outliers is the studentized residual. To work with this statistic in the app, the user simply clicks the button corresponding to 'Studentized residuals' for the measure of interest. They then input a value for the desired cutoff, which represents the value such that any observation that takes on a statistic value (e.g., studentized residual) of that number or larger is identified in the scatterplot, along with the value. Figure 5 displays the output after inputting a value of two for the cutoff, with purple line segments drawn from individual points to the regression line (and studentized residual values added to the plot) when points are deemed regression outliers based on the studentized residual cutoff of two specified by the user.



Step 1

Select the variables of interest. Then click the 'Make Plot' button.

Y Variable: Price

X Variable: House.Size

Optional: Customize the axis labels, if desired.

X-axis Label: House Size (square feet)

Y-axis Label: Selling Price ($)

Step 2

Click on a point in the scatterplot below, and drag it to view how the regression output (located at the bottom of the app) changes, if desired.

New Output

```
Y-int            5.64e+04
Slope                84.9
Correlation        0.7648
R^2                0.5849
SD of Residuals  6.974e+04
Slope p-value    1.069e-38
```

Original Output

```
Y-int            5.64e+04
Slope                84.9
Correlation        0.7648
R^2                0.5849
SD of Residuals  6.974e+04
Slope p-value    1.069e-38
```

Step 3

View regression outliers, high leverage points, or influential observations by selecting the measure of interest. For information on each measure, see the 'Help' tab.

Measure of Interest

○ None
○ SD of residuals lines
● Studentized residuals
○ Cooks distance
○ Leverage values
○ DFFITS
○ DFBETAS

Additional Inputs, when applicable

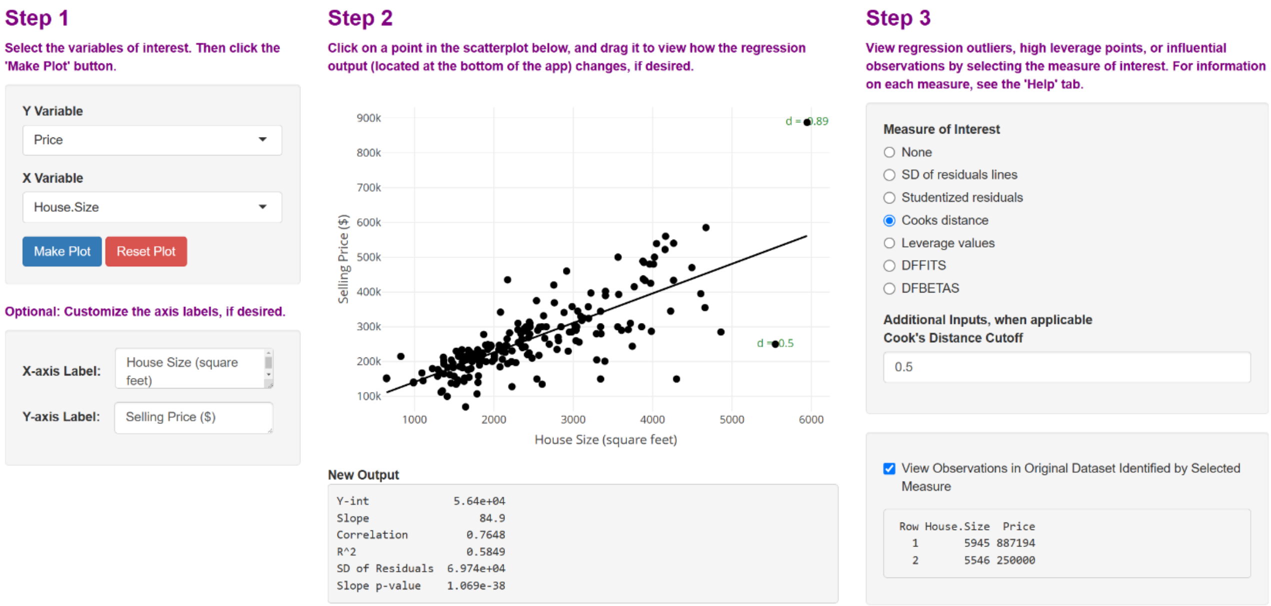
Studentized Residuals Cutoff: 2

☑ View Observations in Original Dataset Identified by Selected Measure

```
Row House.Size  Price
  1       5945 887194
  2       5546 250000
  4       4860 285000
  9       4301 150000
 13       4163 560000
 15       4049 539000
 32       3564 500000
 36       3398 201000
 39       3344 150000
 60       2916 460000
 72       2608 135000
105       2173 435000
```

Figure 5. When the user clicks the 'Studentized residuals' button and inputs a cutoff (here, two) for detecting regression outliers using this statistic, the observations with studentized residual values equal to or above the cutoff can be identified graphically in the scatterplot, with the respective studentized residual values displayed in purple. The row numbers corresponding to those points are printed in the bottom-right-hand corner of the app. In this case, with the specified cutoff of two, there are 12 potential regression outliers. Of course, this cutoff is quite conservative, and there would be fewer outliers if a cutoff of three were used instead.

When assessing whether an observation is influential, three typical statistics one can use are Cook's distance, DFFITS, and DFBETAS. To work with Cook's distance in the app, the user simply needs to click the 'Cooks distance' button and input a cutoff value. Figure 6 displays the output for the housing example based on a conservative Cook's distance cutoff of 0.5 (where a smaller cutoff



Step 1

Select the variables of interest. Then click the 'Make Plot' button.

Y Variable: Price

X Variable: House.Size

Optional: Customize the axis labels, if desired.

X-axis Label: House Size (square feet)

Y-axis Label: Selling Price ($)

Step 2

Click on a point in the scatterplot below, and drag it to view how the regression output (located at the bottom of the app) changes, if desired.

New Output

```
Y-int              5.64e+04
Slope                  84.9
Correlation          0.7648
R^2                  0.5849
SD of Residuals   6.974e+04
Slope p-value     1.069e-38
```

Step 3

View regression outliers, high leverage points, or influential observations by selecting the measure of interest. For information on each measure, see the 'Help' tab.

Measure of Interest

○ None
○ SD of residuals lines
○ Studentized residuals
● Cooks distance
○ Leverage values
○ DFFITS
○ DFBETAS

Additional Inputs, when applicable

Cook's Distance Cutoff: 0.5

☑ View Observations in Original Dataset Identified by Selected Measure

```
Row House.Size  Price
  1       5945 887194
  2       5546 250000
```

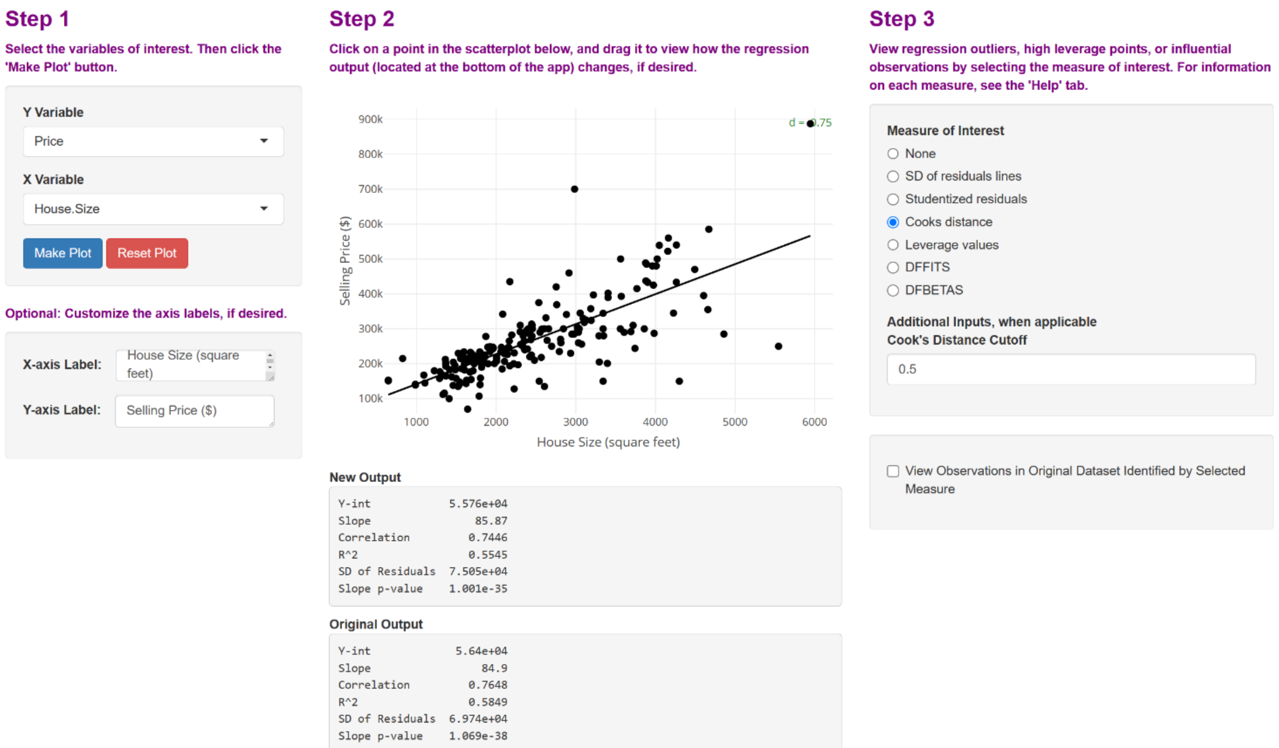
Figure 6. After the user clicks the ‘Cooks distance’ button and inputs a Cook’s distance cutoff (0.5 here), any observations with a Cook’s distance of the specified cutoff or larger is potentially influential. Here, two observations might be influential based on the user’s selections.

would be less conservative). Based on the output, two observations are classified as potential influential observations based on Cook’s distance with a cutoff of 0.5.

DFFITS and DFBETAS are also built into the app when exploring potential influential observations, as is leverage when assessing whether any points have high leverage. While the author has chosen to leave out screenshots of the app when working with these statistics, users should note that the steps are similar to those when working with Cook’s distance. The user simply selects the button corresponding to the statistic of interest and inputs a cutoff, after which the app displays (in the scatterplot) the statistic values for the observations that meet the criteria. Users also have the option to view the row numbers of the identified observations, in the same way as before.

In addition to providing users with various options for detecting potential regression outliers, high leverage points, and influential points, the app enables users to interactively explore these types of observations. Users can do so by clicking and dragging individual points in the scatterplot. This method of exploration is aimed at helping statistics students better understand the differences among regression outliers, high leverage points, and influential points. When moving a point in the scatterplot, users can view the potential effect of individual observations on the regression output. To enable users to assess the potential impact of moving a point on regression statistics (e.g., the regression coefficients or r-squared), both the original regression output and the new regression output (i.e., the output after moving a point) are displayed, as are any updated statistic values (e.g., studentized residuals or Cook’s distance).

Returning to the housing example, the focus moving forward will be on the point located at coordinates of approximately (3,000, 350,000), as seen in Figures 3-6. Suppose the user moves that point to coordinates of approximately (3,000, 700,000), as Figure 7 shows. The new point is a potential regression outlier due to its location outside of the overall trend of the data in the vertical direction, but it is not a high leverage point because its location is not extreme in the horizontal direction relative to the other data. Since the point does not have high leverage, it is not influential. In fact, a comparison of the original and new regression output found below the scatterplot supports the lack of influence of this regression outlier, as the regression statistics across the two sets of output are quite similar. Additionally, the Cook’s distance for that point is under 0.5 (since it is not identified in the plot as having a Cook’s distance of at least 0.5), adding even more support to the point’s lack of influence.



Step 1

Select the variables of interest. Then click the 'Make Plot' button.

Y Variable: Price

X Variable: House.Size

Make Plot | Reset Plot

Optional: Customize the axis labels, if desired.

X-axis Label: House Size (square feet)

Y-axis Label: Selling Price ($)

Step 2

Click on a point in the scatterplot below, and drag it to view how the regression output (located at the bottom of the app) changes, if desired.

New Output

```
Y-int             5.576e+04
Slope                 85.87
Correlation          0.7446
R^2                  0.5545
SD of Residuals   7.505e+04
Slope p-value     1.001e-35
```

Original Output

```
Y-int              5.64e+04
Slope                  84.9
Correlation          0.7648
R^2                  0.5849
SD of Residuals   6.974e+04
Slope p-value     1.069e-38
```

Step 3

View regression outliers, high leverage points, or influential observations by selecting the measure of interest. For information on each measure, see the 'Help' tab.

Measure of Interest

○ None
○ SD of residuals lines
○ Studentized residuals
● Cooks distance
○ Leverage values
○ DFFITS
○ DFBETAS

Additional Inputs, when applicable

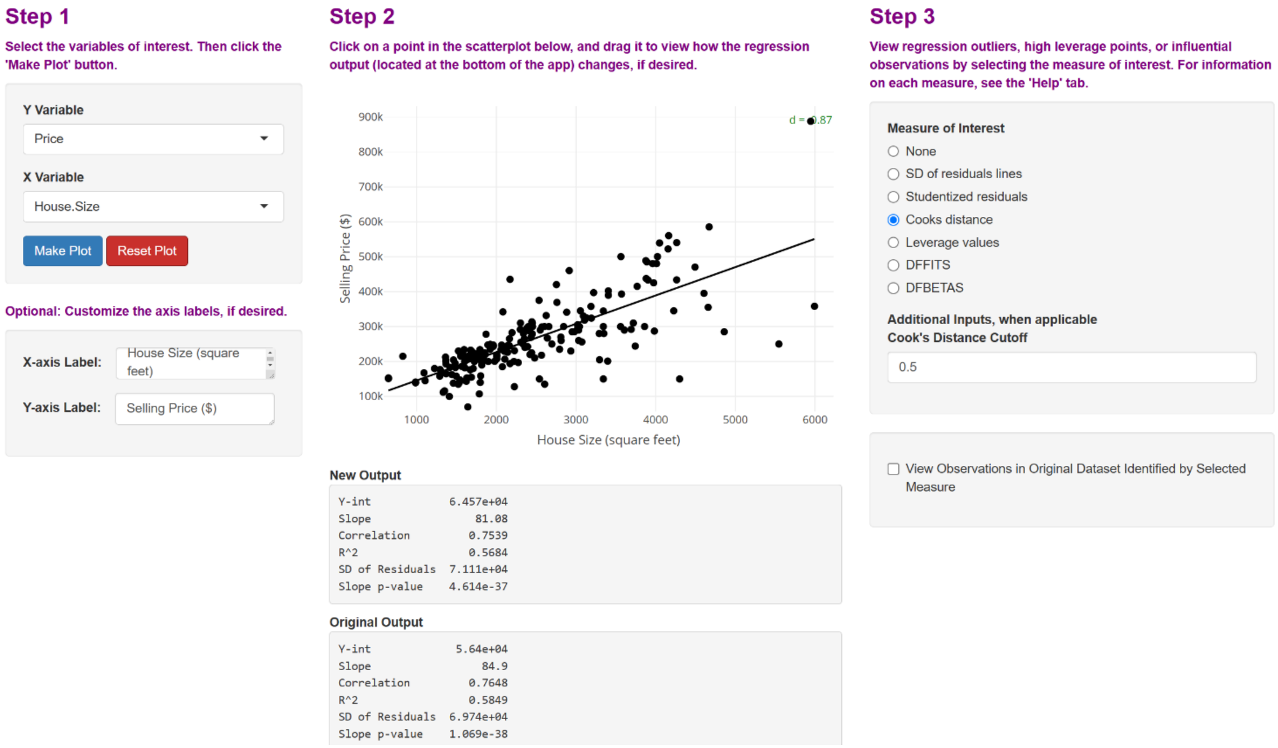
Cook's Distance Cutoff: 0.5

□ View Observations in Original Dataset Identified by Selected Measure

Figure 7. In the scatterplot, the point originally at coordinates of approximately (3,000, 350,000) has been moved to coordinates of approximately (3,000, 700,000). After the move, the observation is a potential regression outlier, but not a high leverage point, and thus is not an influential observation (which is also confirmed by the observation’s small Cook’s distance).

After resetting the plot by clicking the red ‘Reset Plot’ button, suppose the user now moves the same point of interest (i.e., the one originally located at coordinates of roughly (3,000, 350,000)) to coordinates of roughly (6,000, 350,000), as seen in Figure 8. The new point is a potential high leverage point due to its extreme horizontal location relative to most of the other data values, but it is not a regression outlier because it is not particularly extreme in the vertical direction. As a result, the point is not influential. A comparison of the original and new regression output found below the scatterplot supports the lack of influence of this point, as the regression statistics across the two sets of output are quite similar. The observation’s Cook’s distance also supports the lack of influence.

After resetting the plot once again by clicking the ‘Reset Plot’ button, suppose the user now moves the same point of interest (i.e., the one originally located at coordinates of roughly (3,000,



Step 1

Select the variables of interest. Then click the 'Make Plot' button.

Y Variable: Price

X Variable: House.Size

Optional: Customize the axis labels, if desired.

X-axis Label: House Size (square feet)

Y-axis Label: Selling Price ($)

Step 2

Click on a point in the scatterplot below, and drag it to view how the regression output (located at the bottom of the app) changes, if desired.

New Output

```
Y-int              6.457e+04
Slope                  81.08
Correlation           0.7539
R^2                   0.5684
SD of Residuals    7.111e+04
Slope p-value      4.614e-37
```

Original Output

```
Y-int               5.64e+04
Slope                   84.9
Correlation           0.7648
R^2                   0.5849
SD of Residuals    6.974e+04
Slope p-value      1.069e-38
```

Step 3

View regression outliers, high leverage points, or influential observations by selecting the measure of interest. For information on each measure, see the 'Help' tab.

Measure of Interest

○ None
○ SD of residuals lines
○ Studentized residuals
● Cooks distance
○ Leverage values
○ DFFITS
○ DFBETAS

Additional Inputs, when applicable

Cook's Distance Cutoff: 0.5

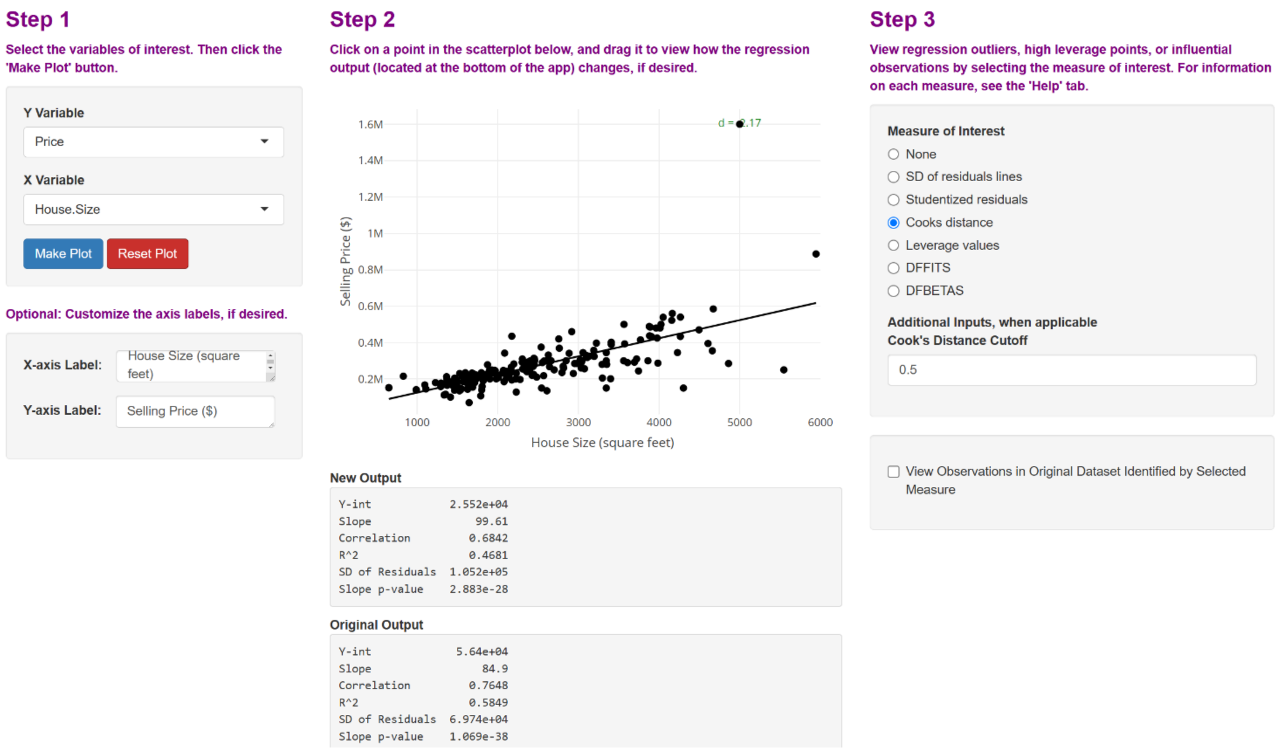
□ View Observations in Original Dataset Identified by Selected Measure

Figure 8. In the scatterplot, the point originally at coordinates of approximately (3,000, 350,000) has been moved to coordinates of approximately (6,000, 350,000). After the move, the observation is a potential high leverage point, but not a regression outlier, and is therefore not an influential observation (which is also confirmed by the observation's small Cook's distance).

350,000)) to coordinates of roughly (5,000, 1,600,000), as seen in Figure 9. The new point is potentially both a regression outlier and a high leverage point due to its falling outside of the overall pattern of the data as well as its extreme horizontal location relative to most of the other data values. As a result, the point is potentially influential. A comparison of the original and new regression output found below the scatterplot indicates the influence of this point, as the regression statistics across the two sets of output are noticeably different. In fact, the observation's Cook's distance of 2.17 is substantially larger than 0.5, which also supports the influence of the observation on the analysis.

While the previous examples of moving a point throughout the scatterplot can be used to explain potential regression outliers, high leverage points, and influential points, users could additionally have the app calculate other measures that were discussed previously in this section (e.g., studentized residuals and leverage). By doing so, instructors and students could then fully tie together all of the ideas discussed in this article.

A final important note regarding moving points in the scatterplot is that the author is in no way suggesting users modify their data when running a regression analysis in practice. The app was developed specifically for pedagogical purposes: to assist (1) faculty when teaching regression outliers, high leverage points, and influential points, and (2) students when learning these important concepts.



Step 1

Select the variables of interest. Then click the 'Make Plot' button.

Y Variable

Price

X Variable

House.Size

Make Plot | Reset Plot

Optional: Customize the axis labels, if desired.

X-axis Label: House Size (square feet)

Y-axis Label: Selling Price ($)

Step 2

Click on a point in the scatterplot below, and drag it to view how the regression output (located at the bottom of the app) changes, if desired.

New Output

```
Y-int           2.552e+04
Slope               99.61
Correlation        0.6842
R^2                0.4681
SD of Residuals 1.052e+05
Slope p-value   2.883e-28
```

Original Output

```
Y-int            5.64e+04
Slope                84.9
Correlation        0.7648
R^2                0.5849
SD of Residuals 6.974e+04
Slope p-value   1.069e-38
```

Step 3

View regression outliers, high leverage points, or influential observations by selecting the measure of interest. For information on each measure, see the 'Help' tab.

Measure of Interest

○ None
○ SD of residuals lines
○ Studentized residuals
● Cooks distance
○ Leverage values
○ DFFITS
○ DFBETAS

Additional Inputs, when applicable

Cook's Distance Cutoff

0.5

□ View Observations in Original Dataset Identified by Selected Measure

Figure 9. In the scatterplot, the point originally at coordinates of approximately (3,000, 350,000) has been moved to coordinates of approximately (5,000, 1,600,000). After the move, the observation is potentially influential, as it is appears to be both a regression outlier and a high leverage point. The observation's large Cook's distance also supports its influence.

4. Assessment of the application

An online survey focused on obtaining feedback from faculty after exploring the app described in this article is now discussed. The survey, which was created by the author and conducted using Qualtrics, was emailed to ten faculty members at Fairfield University, a comprehensive university with more than 6,000 undergraduate and graduate students, during the Fall 2024 semester. All ten of the faculty members regularly teach statistics courses that cover regression. The survey was exempt from review by the Institutional Review Board at Fairfield University and included a required question asking if participants give their consent for their responses to be used anonymously for the study. Of the ten faculty contacted, four responded to the survey, with each of the four giving their consent for their responses to be used anonymously for the study.

While the full survey can be found in the Appendix, the statements and questions contained within it focused on the accessibility, ease of use, and usefulness of the app, with Likert-scale response options provided for each of the statements and open-ended text box responses provided for the questions. The responses to each of the Likert-scale statements are summarized in Table 2. For each statement, 100% of the respondents provided positive responses about the app – consisting of either "Agree" or "Strongly Agree" for the positively worded statements (with a single exception of "Neutral" for one statement), and either "Disagree" or "Strongly Disagree" for the negatively worded statement.

Table 2. Faculty responses to Likert-scale statements about the *Influential Points* application.

Statement: Main Idea	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
App easy to access	0	0	0	1	3
App difficult to use	2	2	0	0	0
App includes all desired implementations of methods for assessing regression outliers and related points	0	0	1	0	3
Dragging individual points in the scatterplot helpful	0	0	0	2	2
Plan on using app when teaching these concepts	0	0	0	0	4

The survey then included open-ended questions asking what the respondents found most useful about the app, whether or not there are any features the app does not provide that the faculty member would find useful, and what they did not like about the app, along with suggestions for improvement. When responding to what they found most useful about the app, faculty said:

- “It is easy and straightforward to use the app and students will surely benefit from getting a visual representation of how outliers in responses and/or predictors affect the linear regression fit.”
- “Availability of [a] ‘missing value’ check box; points are easy to move and manipulate; stores the original output so making comparisons becomes easier; allows me to use different metrics for high leverage/[influential] points.”
- “The ability to quickly switch between methods to identify high leverage points.”
- “Ease of use, and the wide variety of tools it has preprogrammed.”

When asked if there are any features the app does not provide that they would find useful, two respondents indicated there are not, while one indicated they would like to see a table below the original regression output that lists the row information for all of the data points that meet the different criteria. Further, when asked what they did not like about the app, none of the faculty had anything negative to say. When asked about anything that could be improved, three did provide suggestions, all of which were minor in nature:

- “Since there are several methods to visualize high leverage points, perhaps a way for a user to access a definition of each method.”
- “The instruction[s] could mention the location of the ‘explore plot’.”
- “I wonder if it was possible to add x-axis and y-axis labels on the plot.”

Additional positive feedback about the app was left in response to the final question, including:

- “It was easy to use. No changes to suggest.”

- “I liked all the aspects of the app, including the fact that the plot can be downloaded as a png.”

In response to the constructive feedback provided by the faculty members, the author made the following changes to the app:

- The addition of a ‘Home’ tab that contains instructions for how to navigate the app.
- The addition of a table, located in the ‘Explore Plot’ tab, that lists the row information for each data point that meets the criteria specified by the user.
- The ability to add custom x-axis and y-axis labels to the scatterplot in the ‘Explore Plot’ tab.
- The addition of a ‘Help’ tab that contains URL links to websites that provide details of the methods implemented.

5. Conclusion

In this article, a free, web-based application (created by the author) for use by statistics instructors and students when covering regression outliers, high leverage points, and influential observations is described. Feedback provided by faculty members who tested the app highlights the ease of access, ease of use, and usefulness of the app. The author’s hope is that the app will be helpful in both faculty instruction and student comprehension of regression outliers, high leverage points, and influential points at all levels of the statistics curriculum and in far-reaching disciplines. Potential future work includes an evaluation of the tool by statistics students, a formal assessment of the effectiveness of the tool in improving student comprehension of the concepts implemented within it, and the development of another app (or an extension of the current app) focused on these types of observations in multiple linear regression settings.

Acknowledgements

The author offers his gratitude to the editor of eJMT, the anonymous reviewers, and his colleagues for their helpful comments that strengthened both this article and the application described within it.

Appendix

The online survey sent to faculty regarding the app can be found below.

1. Do you permit the use of your responses in a study that will be submitted for publication in a journal? Note that no identifying information will be included, and your responses will remain anonymous. [required question]

   a. Yes, I give my consent for my responses to be used anonymously for the study as described.
   b. No, I do not give my consent for my responses to be used anonymously for the study as described.

Please respond to the following statements and questions about the Influential Points application, making sure to read each carefully.

2. I found the app easy to access.

   a. Strongly Disagree
   b. Disagree
   c. Neutral
   d. Agree
   e. Strongly Agree

3. I found the app difficult to use.

   a. Strongly Disagree
   b. Disagree
   c. Neutral
   d. Agree
   e. Strongly Agree

4. The app includes implementations of all of the methods I teach for determining whether or not a point is a regression outlier, high leverage point, and/or influential point.

   a. Strongly Disagree
   b. Disagree
   c. Neutral
   d. Agree
   e. Strongly Agree

5. I found dragging individual points in the scatterplot helpful for visualizing the potential impact of regression outliers, high leverage points, and influential points on a regression analysis.

   a. Strongly Disagree
   b. Disagree
   c. Neutral
   d. Agree
   e. Strongly Agree

6. I plan on using the app when I teach regression outliers, high leverage points, and/or influential points.

   a. Strongly Disagree
   b. Disagree
   c. Neutral
   d. Agree
   e. Strongly Agree

7. What did you find most useful about the app?

   [text response]

8. Are there any features that the app does **not** provide that you would find useful when you teach regression outliers, high leverage points, and/or influential points?

   [text response]

9. Overall, what did you **not** like about the app? Is there anything you think could be improved? Please explain briefly.

   [text response]

References

[1] Biehler, R., Frischemeier, D., Reading, C., and Shaughnessy, M. (2018). Reasoning About Data In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), International Handbook of Research in Statistics Education (pp. 139-192): Springer International.

[2] Bobek, E., & Tversky, B. (2016). “Creating Visual Explanations Improves Learning,” *Cognitive Research: Principles and Implications*, 1, 1-14.

[3] Chan, C., Leeper, T., Becker, J., & Schoch, D. (2023). *rio: A Swiss-Army Knife for Data File I/O*. R package version 1.2.2.

[4] Chance, B., Ben-Zvi, D., Garfield, J., & Medina, E. (2007). “The Role of Technology in Improving Student Learning of Statistics,” *Technology Innovations in Statistics Education*, 1:1, 1-26.

[5] Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). *shiny: Web Application Framework for R*. R package version 1.9.1.

[6] Garfield, J., & Ben-Zvi, D. (2007). How Students Learn Statistics Revisited: A Current Review of Research on Teaching and Learning, *International Statistical Review*, 75:3, 372-396.

[7] Garfield, J., & Everson, M. (2009). Preparing Teachers of Statistics: A Graduate Course for Future Teachers, *Journal of Statistics Education*, 17:2.

[8] JASP Team (2024). JASP. Version 0.19.1. Available at https://jasp-stats.org/.

[9] McNamara, A. (2018). Key Attributes of a Modern Statistical Computing Tool, *The American Statistician*, 1-30.

[10] Neumann, D., Hood, M., & Neumann, M. (2013). Using Real-life Data When Teaching Statistics: Student Perceptions of This Strategy in an Introductory Statistics Course, *Statistics Education Research Journal*, 12:2, 59-70.

[11] Pea, R. D. (1987). “Cognitive Technologies for Mathematics Education,” In A. Schoenfeld, *Cognitive Science and Mathematics Education* (pp. 89-122). Erlbaum.

[12] Pearson Education (2024). *StatCrunch*. Pearson Education, London, UK. Available at https://www.statcrunch.com.

[13] R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[14] Repenning, A., Webb, D., & Ioannidou, A. (2010). “Scalable game design and the development of a checklist for getting computational thinking into public schools.” *SIGCSE’10*, https://dl.acm.org/citation.cfm?id=1734357.

[15] Rumsey, D. (2002). Statistical Literacy as a Goal for Introductory Statistics Courses, *Journal of Statistics Education*, 10:3.

[16] Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC, Florida.

[17] Soflytics Corp. (2024). *Rguroo*, Encino, CA. Available at https://www.rguroo.com.

[18] Singer, J., & Willett, J. (1990). Improving the Teaching of Applied Statistics: Putting the Data Back into Data Analysis, *The American Statistician*, 44:3, 223-230.