Simulationsstudie Statistik für Fortgeschrittene

Wie reagieren OLS und LAD Schätzer auf Ausreißer in Daten?

Alessio Negrini, Peter Fabisch und Katharina Jacob

Das Ziel unserer Simulationsstudie war es, zu untersuchen wie der Ordinary Least Squares (OLS) Schätzer und der Least absolute deviations (LAD) Schätzer auf Ausreißer in den Daten reagieren, und wie robust beide Schätzer im Vergleich zueinander sind. Dazu haben wir in einem Jupyter Notebook in einen Datensatz zusätzliche Ausreißer künstlich hinzugefügt, um vergleichen zu können wie sich die Ergebnisse des OLS- und LAD-Schätzers abhängig von den Ausreißern im Datensatz verändern.

Für unsere Studie haben wir einen Datensatz mit 235 Datenobjekte verwendet, die das monatliche Einkommen sowie die monatlichen Ausgaben für Lebensmittel von jeweils einem belgischen Haushalt im 19. Jahrhundert beschreiben. Um unsere Ergebnisse zu validieren, haben wir die Studie außerdem mit einem größeren Datensatz durchgeführt, der die CO2-Werte auf Hawaii in einem wöchentlichen Abstand dokumentiert hat.

Abb. 1: Scatterplot des Datensatzes

In unserer Studie haben wir den OLS und LAD Schätzer verwendet, um die Beziehung zwischen den monatlichen Ausgaben für Lebensmittel eines Haushaltes als abhängige Variable und dem monatlichen Haushaltseinkommen als unabhängige Variable in einer linearen Regression zu schätzen.

Y_i: montliche Ausgaben für Lebensmittel von Haushalt i, $i = 1, \dots, 235$
X_i: Montaliches Einkommen von Haushalt i , $i = 1, \dots, 235$

$$\beta = (\beta_0, \beta_1), \qquad X = \begin{pmatrix} 1, & X_1 \\ \dots & \\ 1, & X_n \end{pmatrix}, e = \begin{pmatrix} e_1 \\ \dots \\ e_n \end{pmatrix}, Y = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix}$$

Stichprobenmodel: $Y = \beta X + e$ Statistisches Modell: $\hat{Y} = \beta X$

Der OLS-Schätzer sucht nach den optimalen Parametern β_0 und β_1 für das Model, welche die Summe der quadratischen Fehler zwischen den geschätzten und den beobachteten Werten der abhängigen Variablen minimieren. Im Gegensatz zum OLS-Schätzer, minimiert der LAD Schätzer die Summe der absoluten Fehler zwischen den geschätzten und den tatsächlichen Werten der abhängigen Variablen. Wie beim OLS Schätzer ist es auch beim LAD Schätzer wichtig, dass die Annahmen der linearen Regression erfüllt sind, um zuverlässige Schätzungen zu erhalten.

Im ersten Schritt unserer Simulation haben eine OLS und LAD Schätzung mit dem ursprünglichen, unveränderten Datensatz durchgeführt, welcher unsere Grundgesamtheit darstellt. Dann haben wir dem Datenset in einem zweiten Schritt künstlich Ausreißer im Umfang von 10% seiner ursprünglichen Größe hinzugefügt. Dazu haben wir uns mit dem gewichteten Interquartilsabstands zwischen dem 25% und 75% Quantil sowie der gewichteten Standardabweichung eine obere und untere Grenze erzeugt, in der die Ausreißer-Haushaltseinkommen liegen sollen:

$$[q_{75} + 1{,}5 * IQR + 3 * std \,;\, q_{75} + 1{,}5 * IQR + 4 * std]$$

Anschließend haben wir die Ausreißer über eine Gleichverteilung daraus gezogen. Der zugehörige Y-Wert (foodexp) wurde mit einer Gleichverteilung aus dem Intervall zwischen 25% und 75% Quantil der Randverteilung (foodexp) gezogen. Unsere Ausreißer sind somit Haushalte, die, gegeben ihre monatlichen Lebensmittelausgaben, ein besonders hohes Monatseinkommen haben. Da die Ausgaben der Haushalte für Lebensmittel ihr Einkommen sinngemäß nicht übersteigen können, haben wir keine Ausreißer mit ungewöhnlich niedrigen Einkommen erzeugt.

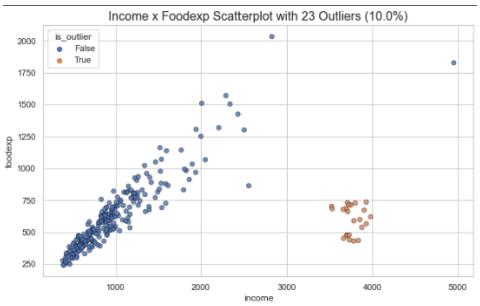

Abb. 2: Scatterplot des augmentierten Datensatzes

In der eigentlichen Simulation haben wir 1000 Stichproben mit jeweils 100 Instanzen aus unserer Grundgesamtheit mit Zurücklegen gezogen, und auf der Stichprobe eine OLS- und eine LAD-Regression durchgeführt. Dieses Vorgehen haben zuerst auf unsere unveränderte, ursprüngliche Grundgesamtheit angewandt und anschließend im selben Umfang auf den Daten der augmentierten Grundgesamtheit mit den künstlich erzeugten Ausreißern wiederholt.

Um die Ergebnisse vergleichen zu können haben, wir in einem Histogramm die Verteilung des Schätzer-Parameters β_1 und des Intercepts über die 1000 Ziehungen dargestellt, und den Mittelwert und die Standardabweichungen der Paramater in dieser Verteilung berechnet.

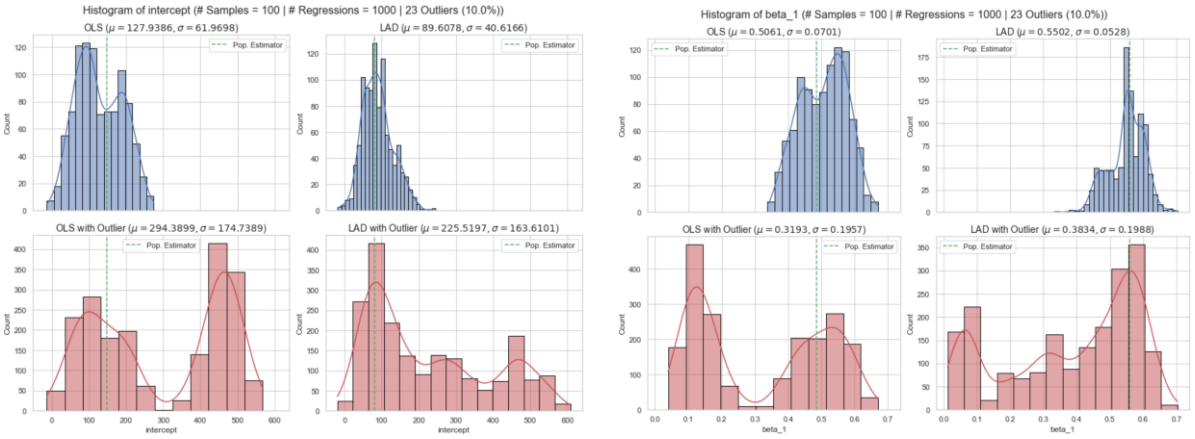

Abb. 3 & 4: Histogramme mit den Lageparametern des Intercepts und von β_1 über die 1000 Regressionen

$$\Delta\mu_{OLS}^{\beta_0} = 166.45, \quad \Delta\mu_{LAD}^{\beta_0} = 135.91, \quad \Delta\sigma_{OLS}^{\beta_0} = 112.77, \quad \Delta\sigma_{LAD}^{\beta_0} = 122.99$$
$$\Delta\mu_{OLS}^{\beta_1} = -0.1868, \quad \Delta\mu_{LAD}^{\beta_1} = -0.1668, \quad \Delta\sigma_{OLS}^{\beta_1} = 0.1256, \quad \Delta\sigma_{LAD}^{\beta_1} = 0.146$$

Bezüglich des Mittelwertes des Intecepts des OLS und des LAD Schätzer zeigt unsere Studie, dass das durchschnittliche Intercept durch das Hinzufügen der Ausreißer deutlich gestiegen ist, beim OLS Schätzer mehr als beim LAD Schätzer. Betrachte man hingegen den Mittelwert des Parameters der unabhängigen Variabel, β_1, kann man in den Ergebnissen unsere Studie eine Verringerung feststellen, auch hier zeigt sich der Effekt deutlicher beim OLS Schätzer. Beim Betrachten der Standardabweichung von β_0 und β_1 lässt sich unserer Studie entnehmen, dass die Standardabweichung der geschätzten Parameter bei beiden Verfahren gestiegen ist. Hier zeigt sich der Effekt allerdings bei LAD deutlicher als bei OLS.

Unsere Beobachtungen zeigen, dass durch das Hinzufügen der Ausreißer ein kleinerer Teil der abhängigen Variable Einkommen durch die unabhängige Variable Lebensmittelausgeben beschrieben wird. Die Ausreißer sorgen dafür, dass weniger Zufall vom Model erklärt werden kann und Vorhersagen ungenauer werden. In unserer Studie zeigt sich dieser Effekt beim OLS Schätzer deutlicher als beim LAD Schätzer, da dort die Verschiebung des Intercepts und von β_1 größer ist. Diese Beobachtungen lassen sich dadurch erklären, dass OLS die quadratischen Abweichungen, und LAD die absoluten Abweichungen zwischen Vorhersagen und tatsächlichen Werten minimiert. Demnach kann ein einziger Ausreißer die Vorhersagen von OLS stärker verzerren, was LAD robuster gegenüber Ausreißern macht, da dort die Ausreißer nicht so stark Gewicht haben.