Research Ideas About Outlier Detection and Robust Regression

Pooya Taherkhani

March 12, 2021

1 Outliers

What are outliers?

To define outliers in the context of regression, consider the design matrix X and the response variable y.

The design matrix X represents n observations[*] $\boldsymbol{x}_i$ each consisting of k measurements.

$$X = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nk} \end{pmatrix}$$

Note that $\boldsymbol{x}_i$ is a vector, such that $\boldsymbol{x}_i^T$ constitutes a row in the design matrix X.

$$\boldsymbol{x}_i^T = \begin{pmatrix} x_{i1} & \dots & x_{ik} \end{pmatrix}$$

Outliers are observations outside the range of the bulk of data, which includes the following.

1. Observations outside the range of X for majority of data points.
   Rows of X are in the k-dimensional space $\mathbb{R}^k$. And thus the range (in the *statistical* sense, not the *analytical*[†]) of X would be a subspace of $\mathbb{R}^k$.
2. Observations outside the range of y for majority of data points.
3. Observations that are outside the range of X and outside the range of y.
4. Observations that are inside the range of X and inside the range of y, yet considered outliers because they are outside the cloud of the majority of data.

Outliers are also called *anomolies*.

[*] In fact $(\boldsymbol{x}_i, y_i)$ is an observation because an observation would *not* be complete without the response variable y_i.

[†] There is actually *not* much difference, if at all, between the statistical and the analytical sense of the word *range*, is there?

1.1 Outliers Revisited

Back to our definition of outliers as observations $\{(x_j, y_j)\}$ that fall outside the bulk of data $\{(x_i, y_i)\}_{i=1}^{n} - \{(x_j, y_j)\}$, we observe that we might be able to offer a single definition that covers all four categories stated above.

1.1.1 Questions

1. How can we formally define a subspace of $\mathbb{R}^{k+1}$ as the *range of the bulk of data*? The closest I have come so far to a formal definition is a subspace such as $\{(x_i, y_i)\}_{i=1}^{n} - \{(x_j, y_j)\}$ where $\{(x_j, y_j)\}$ is the set of outliers.
   I just noticed that this may be a *cyclic* definition! It seems that I am using outliers to define outliers!
2. How can such a definition be useful for our purposes? Can't we just get by, using only the four catergories of outliers outlined above?
3. Shouldn't we take into account the *ratio* of points that we consider as outliers to define *bulk*, and the distance of those points from the bulk to define being *outside* the bulk?
4. Instead of *ratio*, we may be able to use distributions. For example normal distribution of the variable or its logarithm.

1.1.2 Findings

1. The *main cloud of data* or the *range of the bulk of data in the* $\mathbb{R}^{k+1}$ *space* can be defined using a *distribution* for the residuals.
   (a) where residuals are the difference between the predicted and the actual values of the response variable.
2. Points with residuals that deviate from the distribustion (remember distribution represents the *bulk* or the *major cloud* of data) are identified as *outliers*.
3. A defintion of outliers based on the distribution of residuals will spare us the need to classify outliers based on x_i or y_i.

Question How do we define the distribution of residuals for normal points?

4. Any presumed distribution is making assumptions about the nature of the residuals. What assumptions are *appropriate* for the distribution of residuals of bulk of data? What assumptions represent the nature of the residuals? Are there any *universal* assumptions that can be made about residuals of *all* sorts of data and models? Can the assumptions be made independent of the data and models? Or do all the assumptions depend on specific of the data and the model?

Question Is there any benefit to assuming a distribution for the residuals of the normal points? Can we do without any presupposed distribution for the residuals of the normal points? Is there any obligation to assume a distribution for the residuals of the normal points?

It seems that the idea of the bulk of the data (and thus the idea of the outliers as the complement of the bulk!) can be captured by the robust regression methods devised by statistitians, such as Huber (look at first chapter of Robust Statistics book(2018)).

[‡]

[‡] General guideline for research: Think in terms of clear simple ideas, then connect the ideas with each other or connect them to ideas of other people published in the literature. Thinking clear and in terms of simple ideas helps you draw a roadmap first, and then fill in the details. Drawing a mental roadmap helps make real progress in research, whereas getting bogged down in details in the initial steps doesn't!

2 Least Squares cannot Detect Outliers

Here is a dataset, with three outliers, and the Least Squares fitting line with and without considering the outliers plotted respectively in blue and red.

Show that the normalized residuals of least squares does not distinguish the outliers from the bulk of the data.

Question How can I demonstrate the effect of outliers on different regression methods in multiple linear regression? Can it be displayed using a matrix plot?

Question Least squares may not be able to detect the outliers, but its normality plot does that! Isn't that so?

Answer The answer is *yes* sometimes, but *not* always. A few outliers are usually detectable using linear regression diagnostics and plots, but a large number of outliers are not easily detected especially if they are in a clustered cloud [1].

3 Outlier Detection

We can think of finding outliers as a clustering problem. Think of data points as a cloud. outliers by the definition form a separate cloud than the majority of data.

Think of data points $\{(x_i, y_i)\}_{i=1}^{n}$ as a cloud (subspace) in the $\mathbb{R}^n$ space.

After detecting the cloud of outliers, find their geometric center (center of gravity), assign a weight to it (to downplay its effect accordintly), and add that weighted point to your data (Is this possible for all types of outliers? it is, but know (describe) what you are doing.) And then just apply linear regression least squares to the new data.

Question What's wrong with this method?

Answer Not much, but these are the disadvantages: it may involve more computation than a robust regression does.

Note that robust regression can already detect outliers and limit their effect all in one step (sort of). But to pinpoint the outliers using established robust methods, we need to take another step to analyze the (normalized or studentized) residuals.

That extra step may cost exactly as much computation as any clustering method that we might use.

```r
n <- 20
B0 <- 20
B1 <- 4
x_out <- c(50, 55, 52)
y_out <- c(100, 105, 97)
set.seed(5)
x <- runif(n, 1, 70)
e <- rnorm(n, 0, 10)

y <- B0 + B1 * x + e
lm_without <- lm(y ~ x)
plot(x, y, las = 1, cex.axis = 1.3, cex.lab = 1.5)
abline(lm_without$coeff, col = "blue", lwd = 2)
x_with <- c(x, x_out)
y_with <- c(y, y_out)
points(x_out, y_out)
lm_with <- lm(y_with ~ x_with)
abline(lm_with$coeff, col = "red")
legend("topleft", c("With outliers", "Without outliers"), col = c("red", "blue"),
    lty = 1, lwd = c(1, 2), title = "Least Squares", cex = 1.25)

x_out <- c(30, 32, 35)
y_out <- c(268, 273, 266)
plot(x, y, las = 1, cex.axis = 1.3, cex.lab = 1.5)
abline(lm_without$coeff, col = "blue", lwd = 2)
x_with <- c(x, x_out)
y_with <- c(y, y_out)
points(x_out, y_out)
lm_with <- lm(y_with ~ x_with)
abline(lm_with$coeff, col = "red")
legend("bottomright", c("With outliers", "Without outliers"), col = c("red", "blue"),
    lty = 1, lwd = c(1, 2), title = "Least Squares", cex = 1.25)
```

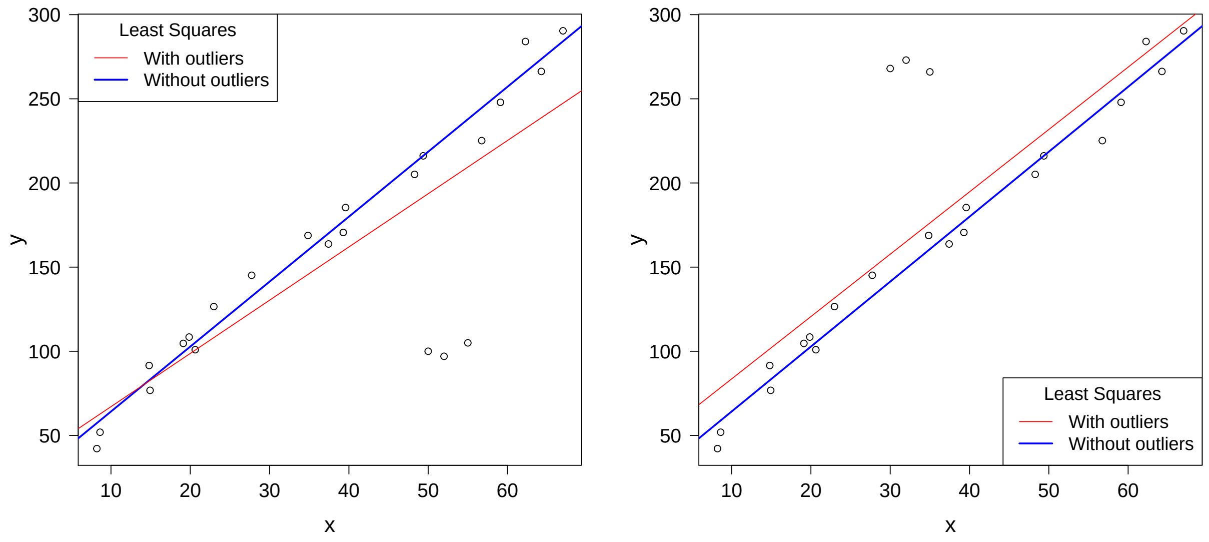

Figure 1: Least Squares *cannot* detect *outliers*.

4 Example

Show through an example in what scenarios the outliers can be detected using linear regression diagnostics and plots.

Figure 1 shows that least squares is sensitive to outliers, and thus, cannot detect them.

Figure 1 was created using the `plot` function among others in R programming language.

5 Roadmap

Following the lead from Adnan et al. [2] and Wisnowsky et al. [1], we develop an intuition from a minial example, then develop a methodology. We then test our methodology on one example. Then test your method on a variety of examples in a systemtic way to generate statistically significant results (this is called Monte Carlo simulation, or just simulation).

6 Data Generation

We generate data from a model (such as $y = \beta_0 + \beta_1 x + \varepsilon$ where $\varepsilon \sim N(0, \sigma)$), then we can apply different modes to the generated data and compare the results. We expect the model based on which the data was generated produces parameters closer to original ones that is $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} << \mathbf{u}$ where $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \dots & \beta_k \end{pmatrix}$ is the vector of regression parameters and $\mathbf{u} = \begin{pmatrix} 1 & \dots & 1 \end{pmatrix}$ is the unit vector.

7 Simulation

Justify your rationale as to why simulation is an appropriate approach (or maybe the only practical way) to compare performance of different regression methods to data containing different types of outliers.

We use simulation under the following conditions.

- Undeterministic nature of the process
- Extremely high number of possible cases
- Determining the outcome analytically may be mathematically intractable.

Question How systematic can we be in designing an experiment that assesses the performance of different regression methods applied to data contaminated by outliers of different forms?

Answer If we cannot systematically count through the number of all possibilities (Bayesian approach), we may need to run random experiments (frequentist approach).

8 Simulation Outline

We define different types of outliers. Then we apply different regression methods to all types of outliers.

Question What is the nondeterministic parameter here?

Answer Maybe

1. number of outliers (or the ratio of outliers to cloud of majority)
2. outliers' relative location (within the same type)
   - location of the cloud of outliers with respect to the location of the cloud of the majority
   - spread of outliers within their cloud

References

[1] J. W. Wisnowski, D. C. Montgomery, and J. R. Simpson, "A comparative analysis of multiple outlier detection procedures in the linear regression model," *Computational Statistics & Data Analysis*, vol. 36, no. 3, pp. 351–382, 2001.

[2] R. Adnan, M. N. Mohamad, and H. Setan, "Multiple outliers detection procedures in linear regression," *MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics*, vol. 19, pp. 29–45, 2003.