Data Analysis, Statistics, Machine Learning

Leland Wilkinson
Adjunct Professor
UIC Computer Science
Chief Scientist
H2O.ai

leland.wilkinson@gmail.com
Anomalies

Anomalies are, literally, lack of a law (*nomos*)

The best-known anomaly is an outlier

This presumes a distribution with tail(s)

All outliers are anomalies, but not all anomalies are outliers

Identifying outliers is not simple

Almost every software system and statistics text gets it wrong

Other anomalies don’t involve distributions

Coding errors in data

Misspellings

Singular events

Often anomalies in residuals are more interesting than the estimated values
Anomalies

Why do we care?

Anomalies may bias statistical estimates

And then again, they might not

You need to worry about influence, not outliers
Anomalies

Why do we care?

Anomalies may bias statistical estimates
- Do NOT drop outliers from a dataset before fitting
- Unless you know why they are outliers
- There are alternatives – robust methods, Winsorizing, trimming, ...

Anomalies may lead to new research ideas
- Give a group of people a battery of psychological tests
- Interview outliers personally

Anomalies may be the needle in the haystack
- Terrorists are rare, extreme
- There may not be enough of them to model their behavior adequately
- Search for anomalies in the general population

Anomalies may lead you to a better model
- You can’t have an anomaly without a model
- Examining anomalies in residuals can help you to modify the model
Anomalies

Outliers

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism” (Hawkins, 1980)

The existing methods in statistics and machine learning packages for detecting outliers based on the mean and standard deviation of a distribution are wrong

That is because, as $n$ increases, critical value of alpha must change in order to prevent false positives

But picking alpha for a given $n$ makes detection of outliers circular

Multivariate outlier detection problem is even harder

Curse of dimensionality means interpoint distances tend toward a constant as $n$ held constant and $p$ heads toward infinity

Graphical methods aren’t much better
Anomalies

Outliers
Don’t bother to Google
You’ll get this...
Anomalies

Outliers

What you will find if you persist
There are two popular tests
Both depend on a normal distribution
Both fail to offer protection for large samples

Grubbs (1950)

Tukey (1977)

The IQR and Outliers

- The IQR is short for “Interquartile Range”
- To calculate IQR, $IQR = Q_3 - Q_1$
- Outliers are calculated using the IQR.
- The rule for outliers is that if a value is outside $1.5(IQR)$ then it is an outlier.
- So, if a value is more than $Q_3 + 1.5(IQR)$ or less than $Q_1 - 1.5(IQR)$ then it is an outlier.

For the two-sided test, the hypothesis of no outliers is rejected if
\[
G > \frac{(N - 1)}{\sqrt{N}} \left( \frac{t^2_{\alpha/2(N-2)}}{N - 2 + t^2_{\alpha/2(N-2)}} \right)
\]

with $t_{\alpha/2(N-2)}$ denoting the critical value of the $t$-distribution with $(N-2)$ degrees of freedom and a significance level of $\alpha/(2N)$.

For the one-sided tests, we use a significance level of $\alpha/N$.

In the above formulas for the critical regions, the Handbook follows the convention that $t_+$ is the upper critical value from the $t$-distribution and $t_-$ is the lower critical value from the $t$-distribution. Note that this is the opposite of what is used in some texts and software programs. In particular, Dataplot uses the opposite convention.
Anomalies

Outliers

Why distance from location (mean, median, ...) is wrong

Remember Hawkins’ definition

“...arouse suspicions that it was generated by a different mechanism”

Wouldn’t you be inclined to say the one on the left is an outlier but not the right?
The two samples have the same mean and standard deviation.
So, the problem boils down to gaps, not distance from center

Dixon (1951)

\[ Q = \text{gap} / \text{range} \]

Tukey-Wainer-Schacht (1978)

\[ z_i = \frac{\sqrt{w_i g_i}}{-\text{midmean}(y)} \], where

\[ w_i = i(n - i) \]
Anomalies

Outliers

Graphical methods

Box plots depend on normal distribution – useless for large $n$

See how many box plot outliers there are for $n = 100,000$?

Letter value box plots (Hofmann, Kafadar, Wickham, 2006) better
Anomalies

Outliers

Graphical methods

Probability plot is one of the best, IF you know the distribution
Anomalies

Outliers

Transformations affect outlier detection
For skewed batches, need to transform before testing for outliers
Anomalies

Skewness and Kurtosis
Use $L$-moments (based on weighted sums)
More robust (no third or fourth powers)

Spikes
Use dot plots
Check for stacks
Signal for Zero Inflated Poisson (ZIP) or other models

Multimodality
Smooth with a kernel
Do bump hunting by computing slope of tangent
Look for more than one bump (mode)
Anomalies

Multivariate Outliers

Mahalanobis Distance is most popular method
OK if you know distribution is multivariate normal
But estimate of covariance matrix can be unreliable when $p$ is large

If so, try computing robust covariances for Mahalanobis Distance
Anomalies

Multivariate Outliers

Principal Components

Plot last few PC’s against each other

As with Mahalanobis Distance, may want to base them on robust covariances
Anomalies

Multivariate Outliers

Minimum Spanning Tree

Compute MST and look for nodes having extremely long edges

![Graph showing PUTOUTRATE and ASSISTRATE with a minimum spanning tree highlighting one node.](image)
Anomalies

Multivariate Outliers

Clustering

1. Choose very large $k$
2. Initialize $k$ centroids
3. Assign every point $y$ to nearest centroid (squared Euclidean distance)
4. Compute within-cluster sum of squares (SSW)
5. Repeat 3 and 4 until SSW does not get noticeably smaller

- On each iteration, use outlier algorithm to decide if a distance to a centroid is beyond cutoff
- If so, leave point out of centroid
- Omitted points are outliers
Anomalies

Multivariate Outliers

Clustering

Didn’t work too well here
Anomalies

Multivariate Outliers

Stahel-Donoho outlyingness

A robust method with high breakdown point

For any real valued vector $y_{p \times 1}$, the measure of outlyingness is

$$r(y, X) = \sup_{a \in S_p} \frac{|a'y - \mu(a'X')|}{\sigma(a'X')}$$

$$S_p = \{ a \in R^p : ||a|| = 1 \}$$

The estimate for $\mu$ is based on the a weighted location estimator

The estimate for $\sigma$ is based on the median absolute deviation (MAD)

The Stahel-Donoho estimator is defined as a weighted mean and covariance, where each observation receives a weight which depends on a measure of its outlyingness. This measure is based on the one-dimensional projection in which the observation is most outlying. The motivation is that every multivariate outlier must be a univariate outlier in some projection.

Computing this is expensive, although one can use sampling to find $a$
Anomalies

Multivariate Anomalies


We characterize a scatterplot (2D point set) with nine measures. We base our measures on three geometric graphs.
Anomalies

Multivariate Anomalies

Scagnostics

**Convex:** area of alpha shape divided by area of convex hull

**Skinny:** ratio of perimeter to area of the alpha shape

**Stringy:** ratio of 2-degree vertices in MST to number of vertices > 1-degree
Anomalies

Multivariate Anomalies

Scagnostics

Skewed: ratio of \((Q_{90} - Q_{50}) / (Q_{90} - Q_{10})\), where quantiles are on MST edge lengths

Clumpy: 1 minus the ratio of the longest edge in the largest runt (blue) to the length of runt-cutting edge (red)

Outlying: proportion of total MST length due to edges adjacent to outliers
Anomalies

Multivariate Anomalies

Scagnostics

**Sparse:** 90th percentile of distribution of edge lengths in MST

**Striated:** proportion of all vertices in the MST that are degree-2 and have a cosine between adjacent edges less than -.75

**Monotonic:** squared Spearman correlation coefficient
Anomalies

Multivariate Anomalies

Scagnostics

Here’s how they distribute in 2D
Anomalies

Multivariate Anomalies

Scagnostics

Original Data Matrix

For each pair of columns in X, we compute 9 measures

Scagnostics Transform

Scagnostics Matrix

\[ q = p(p-1)/2 \]

Copyright © 2016 Leland Wilkinson
Anomalies

Multivariate Anomalies

Detecting outlying scatterplots by cluster analyzing scagnostics matrix

Compute scagnostics matrix and then cluster it

Use cluster outlier method to detect outlying scatterplots

Notice the plot in the upper left is an outlier even though it looks bivariate normal
Anomalies

Multivariate Anomalies

Scagnostics

Ladder of powers transformations reveal different scagnostics under different transformations (Dang & Wilkinson, 2014)
Anomalies

Multivariate Outliers

Outliers in tables

Fit a Poisson (log-linear) model and look at residuals

Bogumił Kamiński, Visualizing tables in ggplot2
Anomalies

Multivariate Outliers

Outliers in tables

Simple chi-square can be used on a two-way table

<table>
<thead>
<tr>
<th>Homosexuality Acceptable</th>
<th>America First</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completely Disagree</td>
<td>Completely Agree</td>
</tr>
<tr>
<td>0.197</td>
<td>2.988</td>
</tr>
<tr>
<td>Mostly Disagree</td>
<td>Mostly Agree</td>
</tr>
<tr>
<td>3.053</td>
<td>5.963</td>
</tr>
<tr>
<td>Mostly Agree</td>
<td></td>
</tr>
<tr>
<td>0.844</td>
<td>-1.420</td>
</tr>
<tr>
<td>Completely Agree</td>
<td></td>
</tr>
<tr>
<td>2.988</td>
<td>-3.275</td>
</tr>
</tbody>
</table>
Anomalies

Inliers

Histograms hide details

Stem-and-leaf and dot plots do not

In this batch, someone rounded some heights of baseball players to nearest inch
Anomalies

Inliers

Detecting duplicates

Pick delta profile distance (Euclidean or other distance metric)
Set delta to zero if you want to detect only exact duplicates
Multivariate sort and flag cases closer than delta
Duplicate cases found in some Iris datasets with this method
Anomalies

Missing Values

A missing value is a value that is not observed.

Rubin (1976) gave missing values a theoretical basis.

Identifying a missing value implies we could measure it under some circumstances.

Missing value categories

- NULL – undefined value (not missing)
- Failure to respond (usually, but not always, missing)
- Refusal to respond (rarely, but sometimes, missing)
- Some other random coding omission

Rubin missing value classes

Relation between a variable and probability of a value being missing:

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

Values must be MAR or MCAR to use Rubin’s Multiple Imputation.
Anomalies

Missing Values

Single imputation (all these methods are invalid)

Hot deck
- randomly select a similar record for imputed value
- reduces uncertainty of estimates

Mean imputation
- replace missing value with mean of variable
- attenuates covariance/correlation estimates

Listwise deletion (standard method in most statistics packages)
- throw out record with any missing values
- reduces power and can introduce bias

Pairwise deletion
- when computing correlations, ignore any case with missing value on either variable
- can induce negative eigenvalues and correlations greater than 1 in absolute value

Regression imputation
- fit regression equation using non-missing cases to predict missing values
- reduces uncertainty of estimates

Copyright © 2016 Leland Wilkinson
Anomalies

Missing Values

Multiple imputation
1. Impute missing values using linear or logistic regression
2. Do this, say, 10 times.
3. Perform the desired analysis on each imputed dataset
4. Average the values of the parameter estimates across the imputed datasets
5. Calculate standard errors of parameters using a formula given by Rubin

The EM Algorithm (for accomplishing step 1 above)
1. Estimate regression coefficients for each missing value
2. Plug estimates into the missing cells
3. Compute covariance matrix on complete data
4. Repeat 1 through 3 until covariance matrix stabilizes
Usually only a few iterations are necessary
Perturb the regression coefficients by a small amount before imputing
Anomalies

Missing Values

Multiple imputation

Can delete up to 50% of values and still get decent estimates

<table>
<thead>
<tr>
<th>Missing Data</th>
<th>Complete Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2. Components loadings</td>
<td>Table 2. Components loadings</td>
</tr>
<tr>
<td>Component(1)</td>
<td>Component(1)</td>
</tr>
<tr>
<td>BABYMORT</td>
<td>BABYMORT</td>
</tr>
<tr>
<td>0.891</td>
<td>0.886</td>
</tr>
<tr>
<td>BIRTH_RT</td>
<td>BIRTH_RT</td>
</tr>
<tr>
<td>0.877</td>
<td>0.876</td>
</tr>
<tr>
<td>HEALTH</td>
<td>EDUC</td>
</tr>
<tr>
<td>-0.865</td>
<td>-0.872</td>
</tr>
<tr>
<td>EDUC</td>
<td>HEALTH</td>
</tr>
<tr>
<td>-0.859</td>
<td>-0.861</td>
</tr>
<tr>
<td>MIL</td>
<td>MIL</td>
</tr>
<tr>
<td>-0.692</td>
<td>-0.695</td>
</tr>
<tr>
<td>DEATH_RT</td>
<td>DEATH_RT</td>
</tr>
<tr>
<td>0.499</td>
<td>0.485</td>
</tr>
</tbody>
</table>

n = 57
References

