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

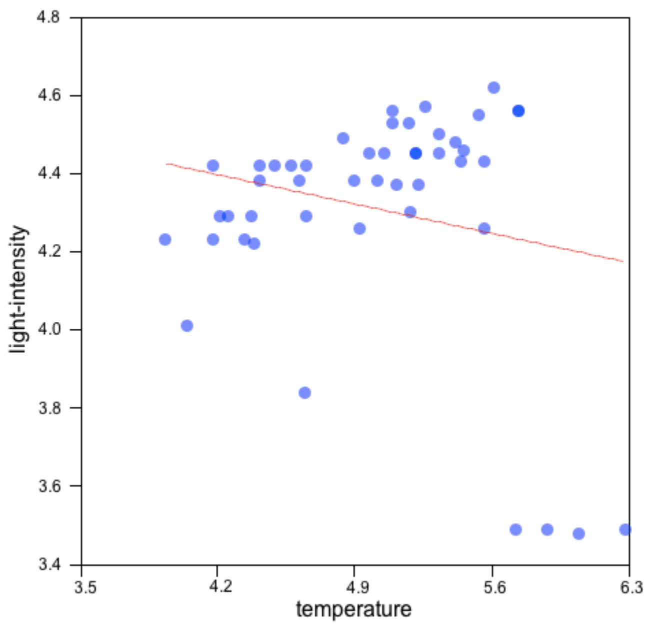
Anomalies

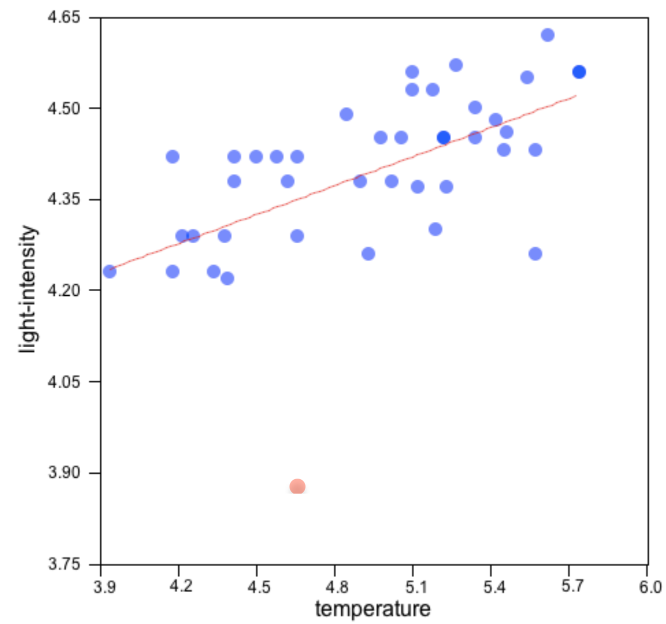
Why do we care?

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

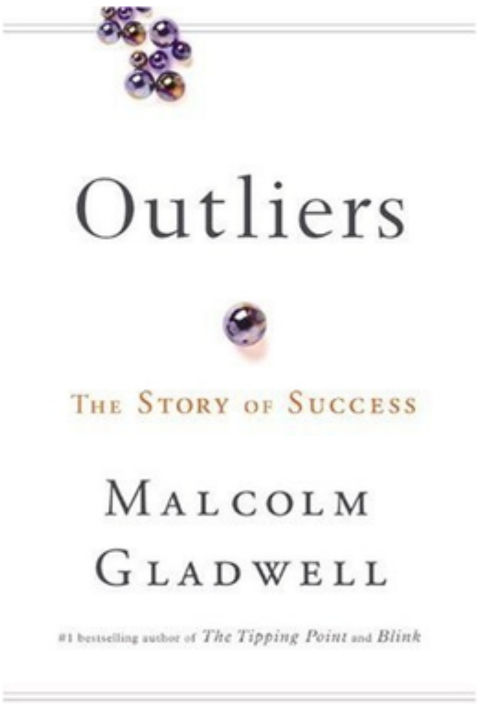
Graphical methods aren't much better

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

Tukey (1977)

Grubbs (1950)

For the two-sided test, the hypothesis of no outliers is rejected if

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/(2N),N-2)}}{N-2+t^2_{(\alpha/(2N),N-2)}}}$$

with $t_{(\alpha/(2N),N-2)}$ 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 α/N.

In the above formulas for the critical regions, the Handbook follows the convention that t_α is the upper critical value from the *t*-distribution and $t_{1-\alpha}$ 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.

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.

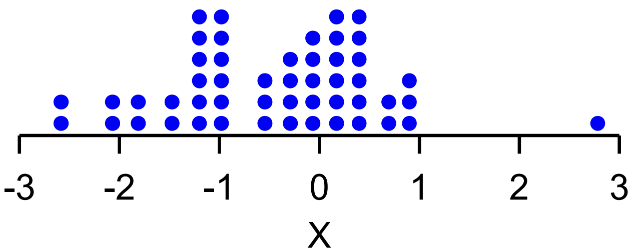
Anomalies

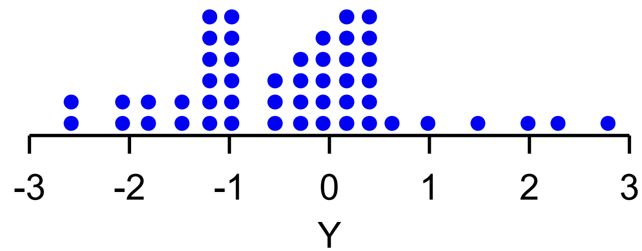
Outliers

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)

$$\begin{aligned} z_i &= \frac{\sqrt{w_i g_i}}{-\text{midmean}(y)} \;, \text{ where} \\ w_i &= i(n-i) \end{aligned}$$

Anomalies

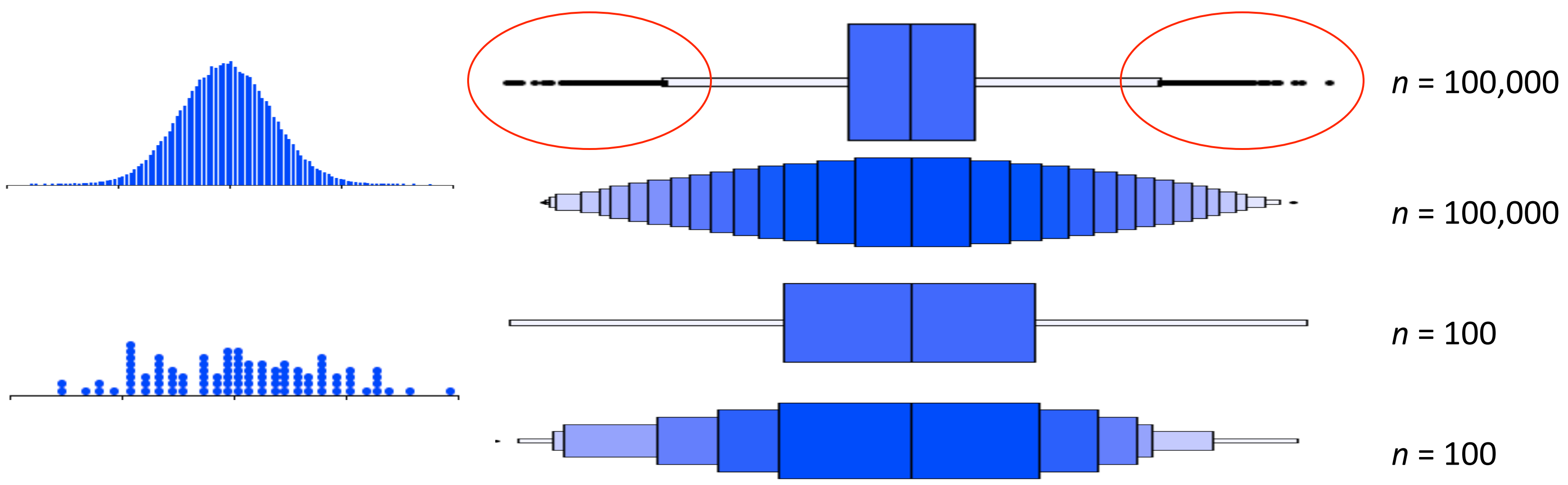
Outliers

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

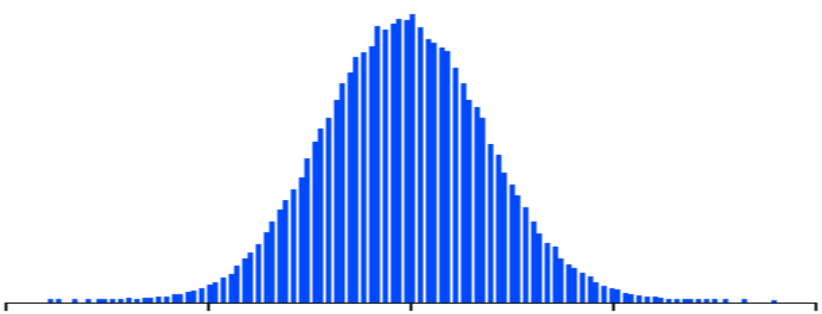


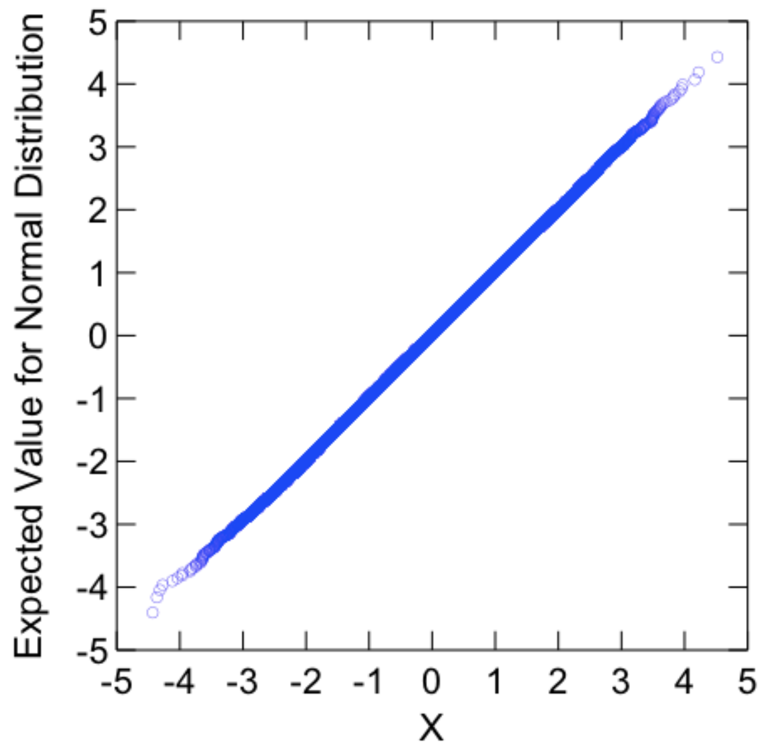
Anomalies

Outliers

Graphical methods

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

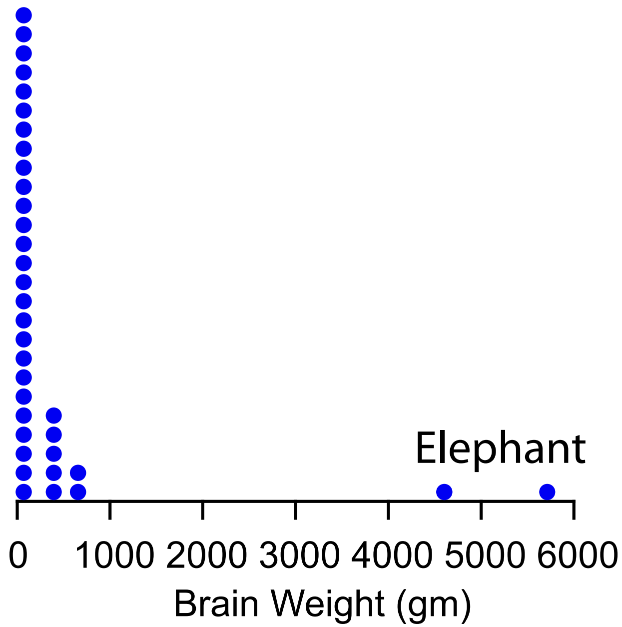


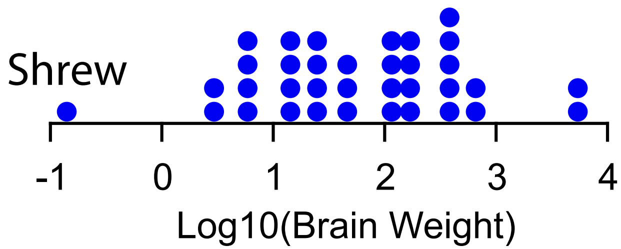


Anomalies

Outliers

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



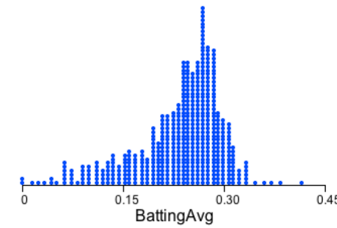


Anomalies

Skewness and Kurtosis

Use *L*-moments (based on weighted sums)

More robust (no third or fourth powers)

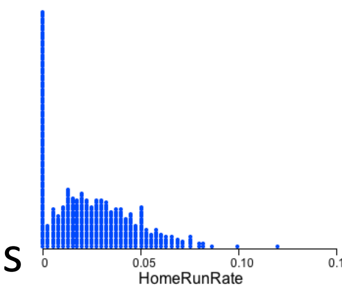


Spikes

Use dot plots

Check for stacks

Signal for Zero Inflated Poisson (ZIP) or other models



Multimodality

Smooth with a kernel

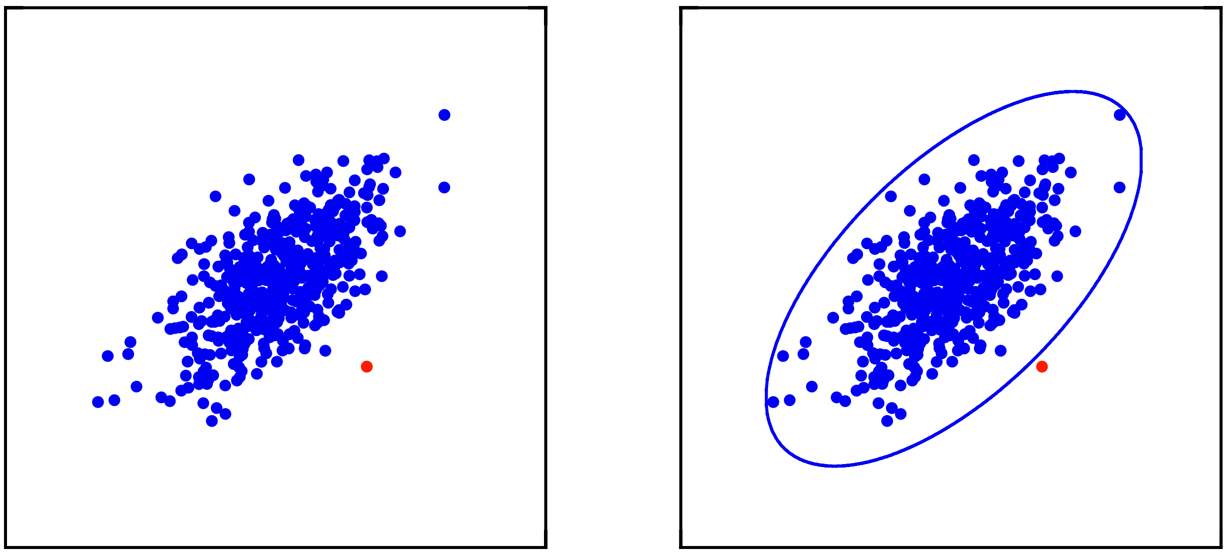
Do bump hunting by computing slope of tangent

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



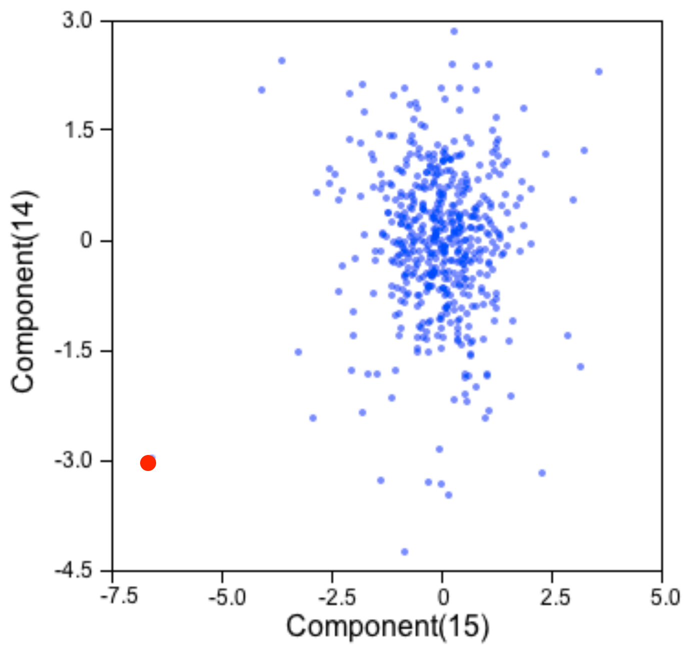
Anomalies

Multivariate Outliers

Principal Components

Plot last few PC's against each other

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

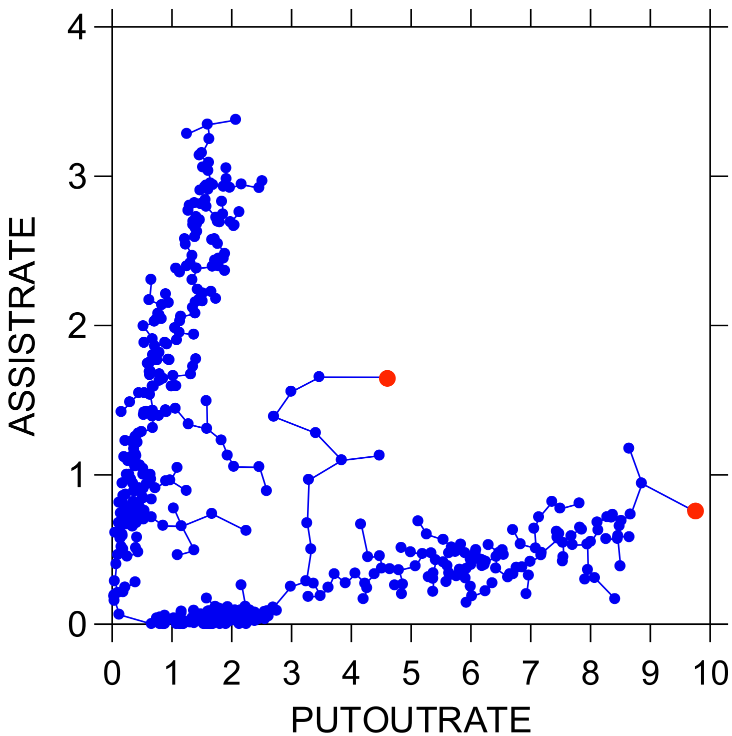


Anomalies

Multivariate Outliers

Minimum Spanning Tree

Compute MST and look for nodes having extremely long edges



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

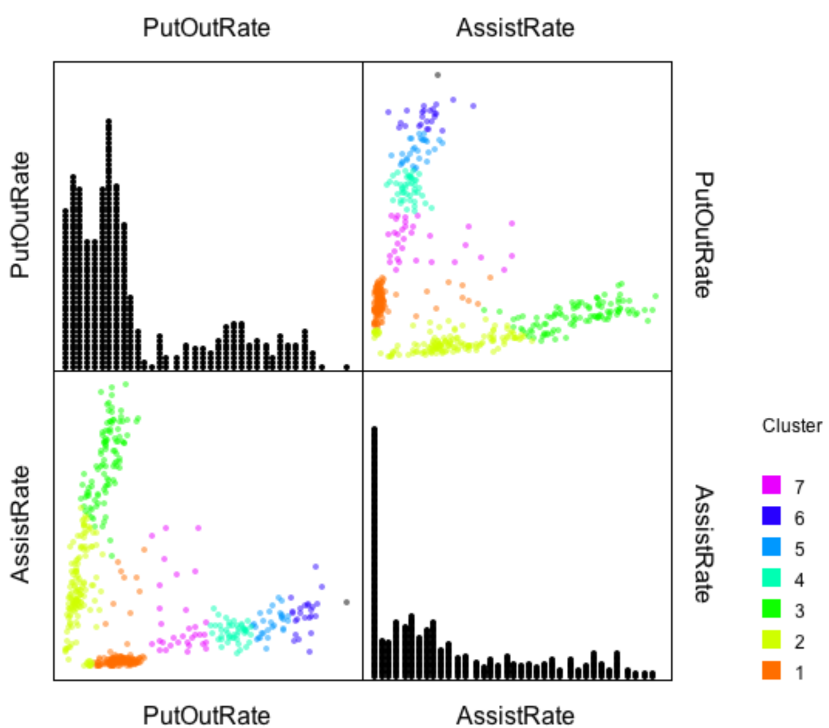
Omitted points are outliers

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 $\mathbf{y}_{p \times 1}$, the measure of outlyingness is

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

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

The estimate for μ is based on the a weighted location estimator

The estimate for σ 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 $\mathbf{a}$

Anomalies

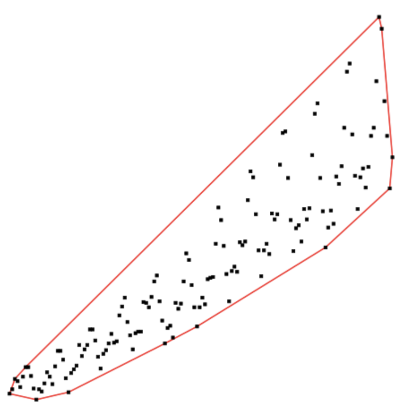
Multivariate Anomalies

Scagnostics (Wilkinson, Anand, Grossman, 2005)

We characterize a scatterplot (2D point set) with nine measures

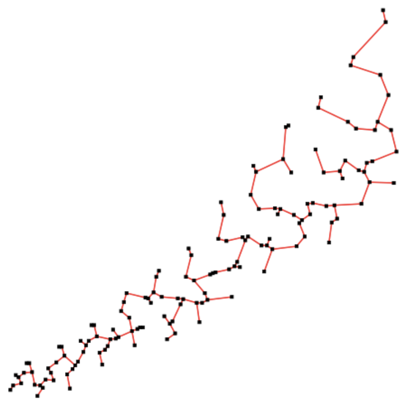
We base our measures on three *geometric graphs*.

Convex Hull



Alpha Shape

Minimum Spanning Tree



Anomalies

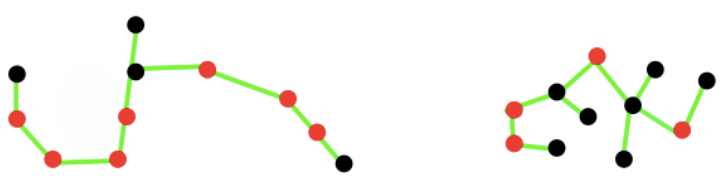
Multivariate Anomalies

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



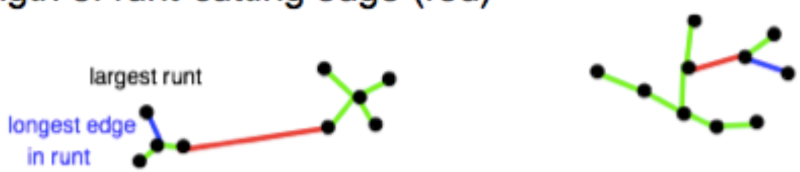
Anomalies

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

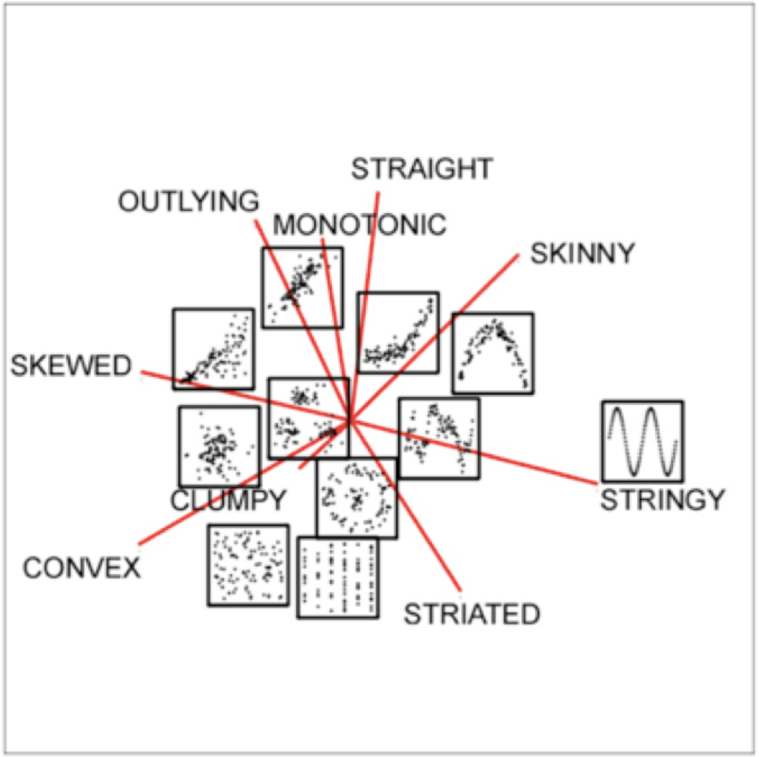
Monotonic: squared Spearman correlation coefficient

Anomalies

Multivariate Anomalies

Scagnostics

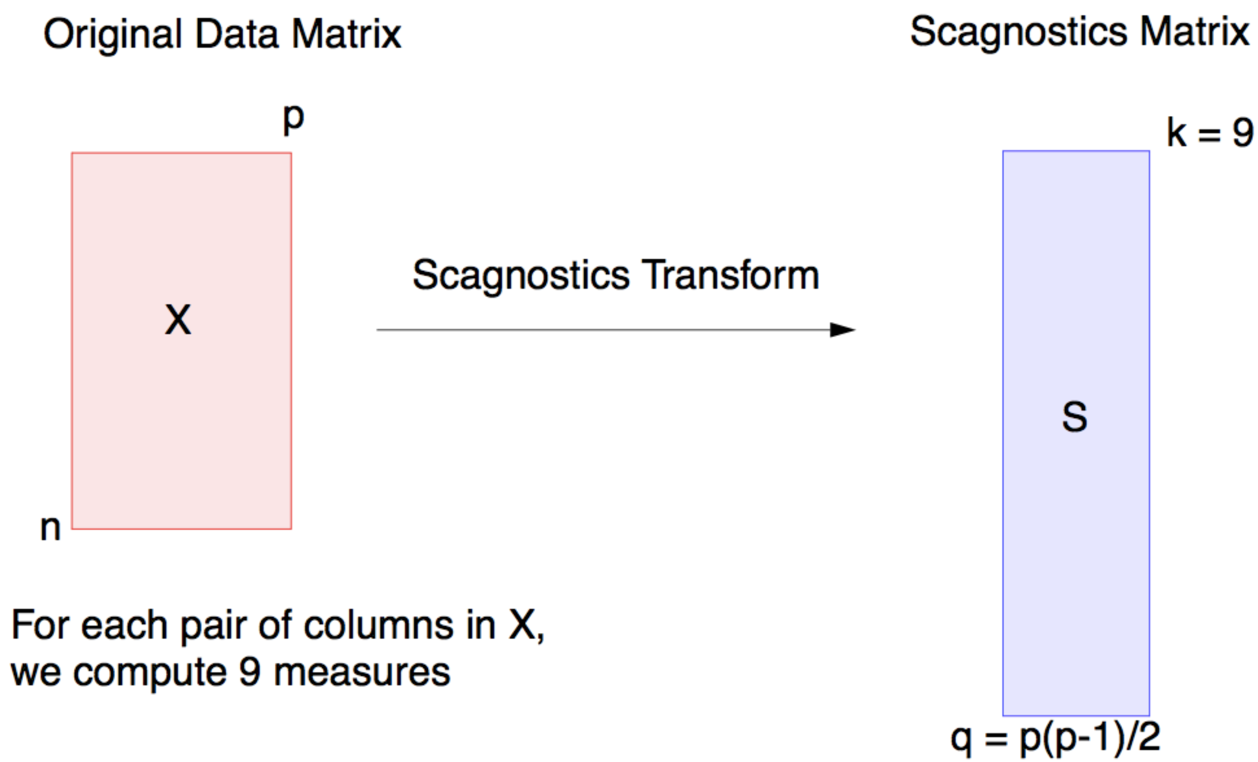
Here's how they distribute in 2D



Anomalies

Multivariate Anomalies

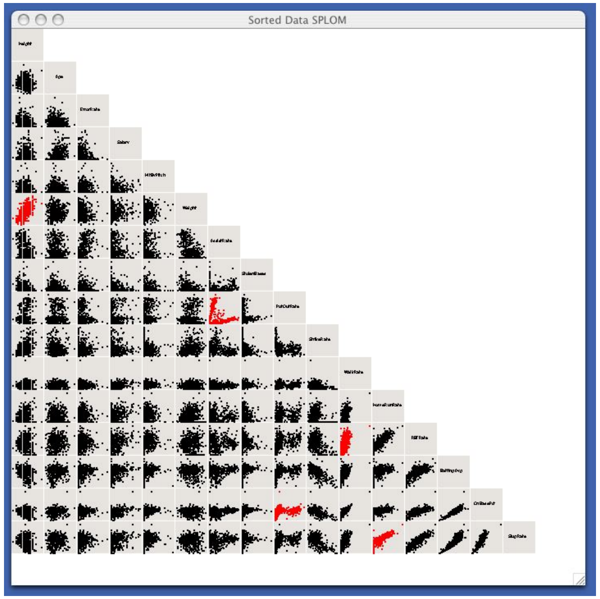
Scagnostics



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

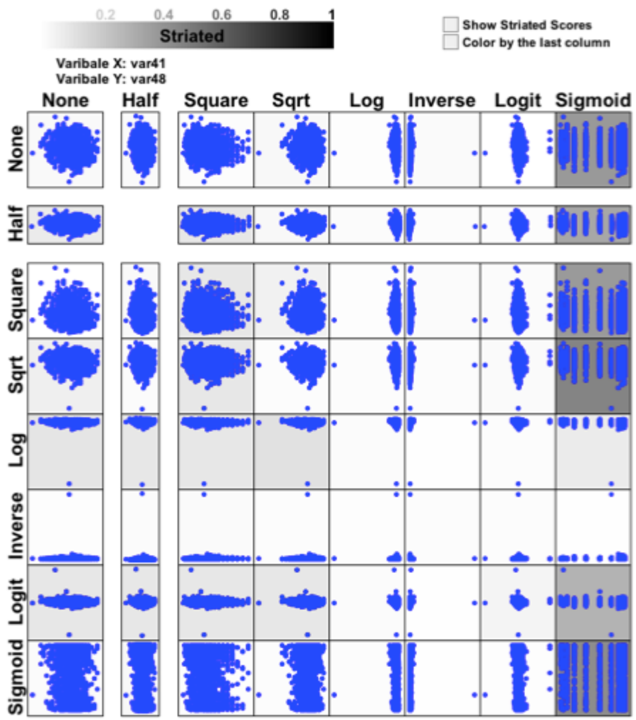


Anomalies

Multivariate Anomalies

Scagnostics

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

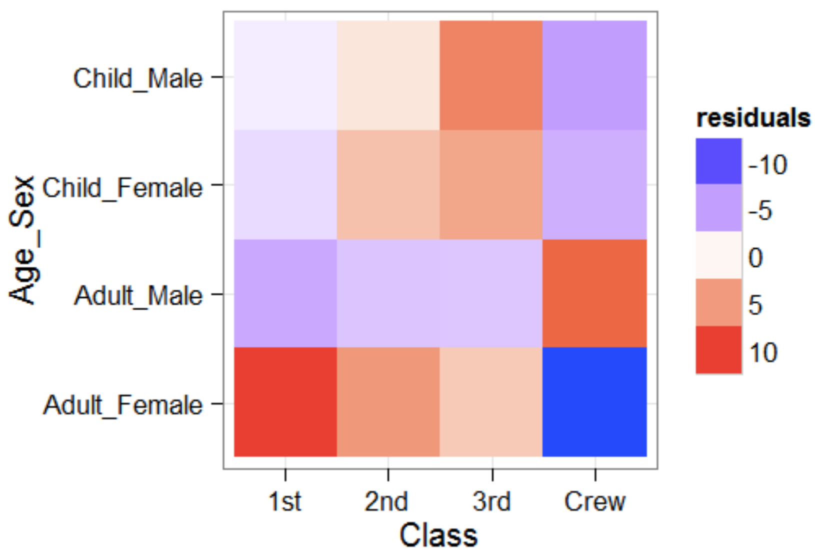


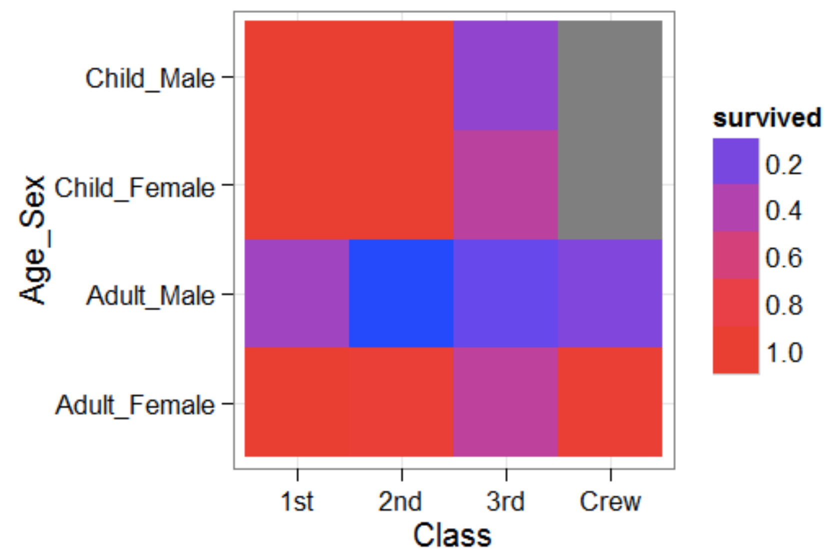
Anomalies

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

America First

Homosexuality Acceptable	Completely Disagree	Mostly Disagree	Mostly Agree	Completely Agree
Completely Disagree	0.197	-2.043	-0.477	2.988
Mostly Disagree	3.053	-2.456	0.605	5.963
Mostly Agree	0.844	1.827	0.514	-1.420
Completely Agree	2.988	0.234	-0.809	-3.275

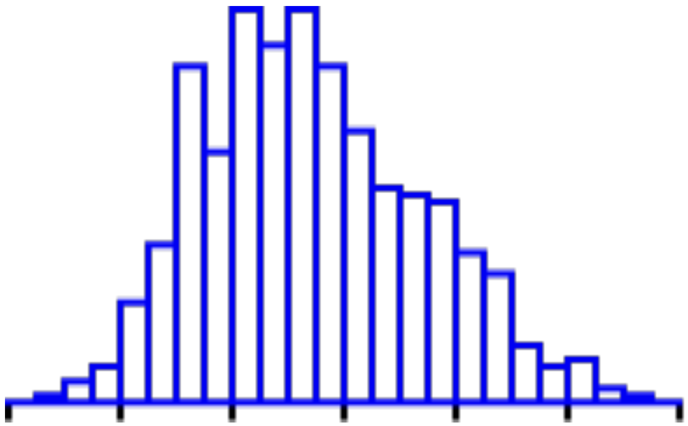
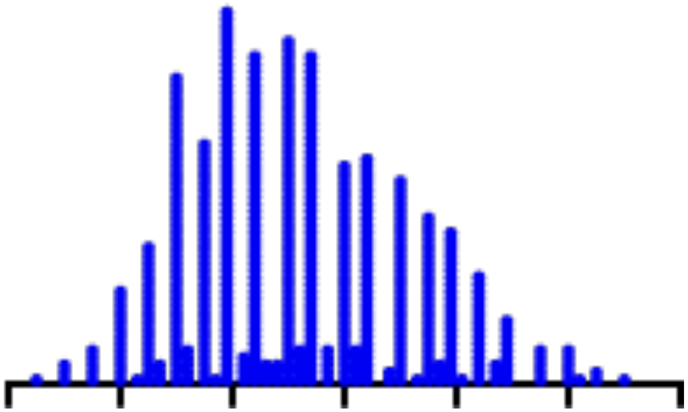
Anomalies

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

368 IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 7, NO. 3, JUNE 1999

Correspondence

Will the *Real* Iris Data Please Stand Up?

James C. Bezdek, James M. Keller, Raghu Krishnapuram, Ludmila I. Kuncheva, and Nikhil R. Pal

***Abstract*—This correspondence points out several published errors in replicates of the well-known Iris data, which was collected in 1935 by Anderson [1], but first published in 1936 by Fisher [2].**

***Index Terms*—Iris data.**

I. Introduction and Conclusions

While preparing Kuncheva and Bezdek [3], these authors discovered that there are (at least) two *distinct* published replicates of the Iris data that have been used as a basis for an unknown number of papers. Subsequently, Bezdek *et al.* [4] discovered two different errors in the version of Iris available through the well-known University of California at Irvine (UCI) machine learning database. Reproduced below, from the preface of Bezdek *et al.* [4] is an account of the problems errors like this cause.

TABLE I
The Iris Data: Fisher [2]

Iris sestosa				Iris versicolor				Iris virginica			
Sepal Leng.	Sepal Width	Petal Leng.	Petal Width	Sepal Leng.	Sepal Width	Petal Leng.	Petal Width	Sepal Leng.	Sepal Width	Petal Leng.	Petal Width
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0

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

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

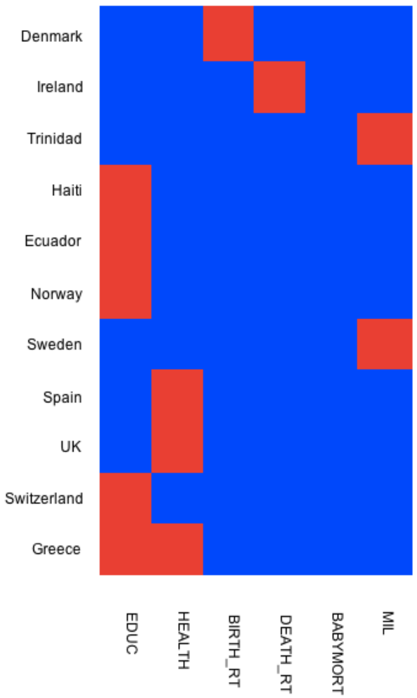
Perturb the regression coefficients by a small amount before imputing

Anomalies

Missing Values

Multiple imputation

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



n = 57

Missing Data

Table 2. Components loadings

	Component(1)
BABYMORT	0.891
BIRTH_RT	0.877
HEALTH	-0.865
EDUC	-0.859
MIL	-0.692
DEATH_RT	0.499

Complete Data

Table 2. Components loadings

	Component(1)
BABYMORT	0.886
BIRTH_RT	0.876
EDUC	-0.872
HEALTH	-0.861
MIL	-0.695
DEATH_RT	0.485

Anomalies

References

Hawkins, D. (1980). *Identification of Outliers*. New York: Chapman and Hall.

Rousseeuw, P.J., and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley & Sons.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.