

Outliers

Daniel Lawson University of Bristol

Lecture 04.2 (v2.0.2)

Why did you exclude
all these responses?



freshspectrum.com

We define outlier as
someone who doesn't
like our program



Signposting

- ▶ How do we identify **Bad Data**? That is, data that is misleading either due to missingness or atypicality.
 - ▶ This is one of the key ways that **Data Science Goes Wrong**.
 - ▶ Most researchers and practitioners do less than they should to understand their data.
- ▶ Today is about outliers. The workshop handles missing data.

Bad Data: Missing and Misleading data

- ▶ The most time-consuming part of any real-world data analysis is **data cleaning**.
- ▶ This takes two main forms:
 - ▶ **Imputing** missing data where possible
 - ▶ **Removing** bad data where necessary
- ▶ It is **vital** that this is handled properly in order to gain appropriate insight from data.

Quality Control: Diagnosing bad data

- ▶ Most of **QC** is about figuring out whether your data are really what you thought they were.
 - ▶ Did you **measure** what you set out to measure?
 - ▶ Are there **systematic effects** that were unexpected?
- ▶ In many disciplines there are well-defined ways to spot issues.

Statistical tools for bad data

- ▶ There are two main tools available:
 1. **Exploratory Data Analysis** (Block 1)
 - ▶ Does it look generally look the way it should?
 - ▶ Methods involve both plots and data summaries
 2. **Outlier Detection**
 - ▶ What specific parts of the data look unusual?
 - ▶ Methods focus on anomaly detection

Key questions to ask

1. Do my data contain important **missingness**?
 - ▶ What aspects of the truth am I not seeing?
 - ▶ How would I know?
 - ▶ What impact could missingness have on my analysis?
2. Do my data containing important **outliers**?
 - ▶ What do we mean by an outlier?
 - ▶ What impact will they have on my subsequent analysis?
 - ▶ What should I do about them?

Anomaly Detection

- ▶ Anomaly detection uses the core methods we have seen throughout.
- ▶ For example, Density estimation (Block 4), cluster analysis (Block 3), regression (Block 2), etc.
- ▶ These models:
 - ▶ provide a baseline measure of **what is Normal**?
 - ▶ Against which **Unusual** is measured.

Measuring “Unusual” with p-values

- ▶ It is straightforward to use any model that can output a p-value as a measure of anomaly.
- ▶ Since a p-value is a Uniform random variable under the null, there is a wide literature available to assess whether the dataset as a whole is anomalous.
- ▶ **The problem:** If there is no plausible null hypothesis,
 - ▶ The data will “look weird” by any statistical measure.

Measuring “Unusual” with descriptive statistics

► **Thresholding:**

- We saw in the “boxplot” that outliers were defined as all observations at least $3/2$ IQR above Q_3 or below Q_1 .
- This comes from reasoning about Normal distributions. . .
- Thresholding can be applied to p-values when they are not interpreted literally.
- Removed values should be investigated to understand why they are unusual.

► Thresholds might be obtained by:

- reference to other datasets,
- theory,
- bootstrapping,
- . . . etc!

Measuring “Unusual” with models

- ▶ Many modelling paradigms **explicitly handle outliers**. Some examples:
- ▶ Regression:
 - ▶ **leverage** of each point (not always the same as outliers)
 - ▶ **Robust regression** methods fit better in the presence of outliers
- ▶ **Density-based** clustering (DBSCAN)
 - ▶ Points in low density regions may be outliers
 - ▶ An empirical p-value can be constructed from the set of points in lower-density regions.
- ▶ **Isolation Forests**
 - ▶ Random Forest-based technique (covered later).
 - ▶ Based on identifying “points that are easy to distinguish with a decision tree”.
- ▶ Many other methods offer $Pr(data|model)$.

Duplicates and sample density

- ▶ **Sample density** obviously affects inference.
 - ▶ The sampling density should reflect the density of the **data to be predicted**.
- ▶ Missing data often makes many records, that **should otherwise be different**, appear the **same**.
 - ▶ This dramatically affects density estimation.
- ▶ One solution is to work only with unique records.
 - ▶ This solves some types of bias but not others, e.g. overrepresentation of particular regions of continuous variables.
 - ▶ No longer a density, but a **plausible region**.

Batch and similar effects

- ▶ Examining associations between features and properties of the data that **should not matter** are a vital tool in Quality Control.
- ▶ Some quantities are known apriori not to affect some feature.
 - ▶ For example, if data are observed in batches, the batch number shouldn't matter.
 - ▶ In regression analyses, minor batch effects can be regressed out (included in the model).
 - ▶ Major batch effects require the data to be discarded or treated specially.
- ▶ As always, **Correlation \neq Causation**.
 - ▶ So observing that e.g. different hospital wards contain systematically different patients isn't a smoking gun for a QC problem.

Robust algorithms

- ▶ Most algorithms have robust alternatives, e.g.
 - ▶ Robust regression, (quantile regression),
 - ▶ Robust clustering,
 - ▶ Robust Kernel Density Estimation,
 - ▶ ... etc. Find one for your problem.
- ▶ Generally, robustness comes at a **cost**:
 - ▶ Increased computational complexity due to e.g. lack of integrability: e.g. Normal kernel replaced by Laplace,
 - ▶ Harder optimisation problem, e.g. more local minima, **non-convex solution**,
 - ▶ Or just not the model you wanted?
- ▶ **Robustness is not a general property** but defined with respect to some class of models.
 - ▶ There are many different “Robust algorithms for X” with different properties.
- ▶ “Too many” outliers will change the model anyway. How many is too many?

Removing 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.” Charu Aggarwal, IBM Research
- ▶ When outliers are detected, what should you do with them?
 - ▶ Switch to a robust algorithm and take the hit?
 - ▶ Remove outliers for the purpose of model building?
 - ▶ Add an “outlier model”, e.g. a larger normal distribution in Gaussian Mixture Modelling?

Reflection

- ▶ How do we know that the class of outliers detected is the “right” ones?
- ▶ Do we expect more outliers in a test dataset?
- ▶ How might we test that an algorithm is the “right kind” of robust?

Signposting

- ▶ Further Reading:

- ▶ "A Survey of Outlier Detection Methodologies" by Victoria Hodge & Jim Austin, Artificial Intelligence Review 22:85–126 (2004).
- ▶ Outlier Analysis by Charu C. Aggarwal. NB: Not freely available.
- ▶ Chapter 10 of The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Friedman, Hastie and Tibshirani) discusses the robustness to outliers for various methods.