



OUTLIERS

The Good and the Bad

Murage Ngatia

June 20, 2012

What are Outliers?

- **Many names:** outliers, anomalies, extreme observations, contaminants, novelties, exceptions, excursions, noise (Barnett and Lewis 1998, Hodge and Austin 2004)
- Grubbs' definition: an outlier is a data point "***that appears to deviate markedly from other members of the sample in which it occurs***" (Grubbs 1969). i.e., outliers are very different from the other data in the sample. Another way of saying this is that outliers come from a different distribution than the rest of the data.
- Sources of outliers
 - Measurement errors (invalid data)
 - Other possible invalid data errors: data logger & transmission problems
 - *Surprising **valid data** that may be more useful than the rest of the data*

Outliers: The Bad!

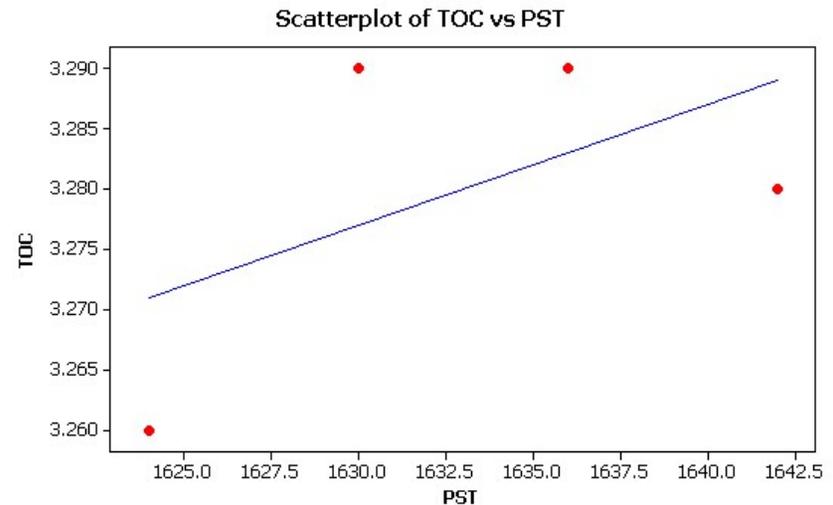
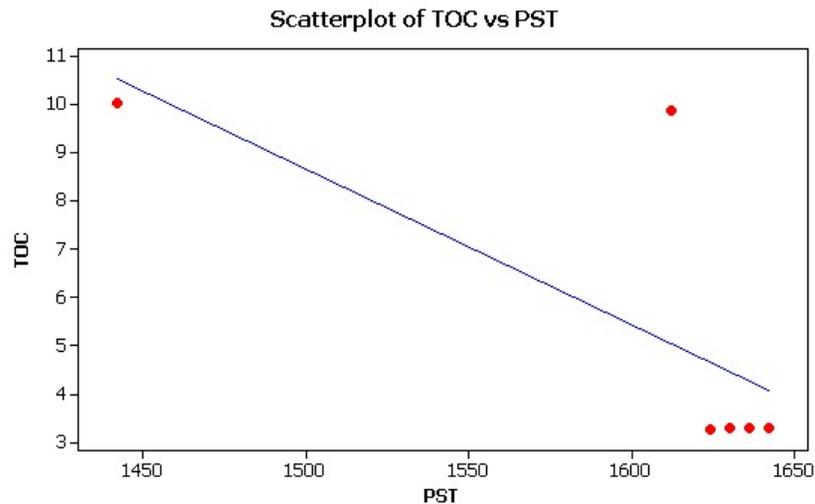
- Significantly **impact: mean, standard deviation, skewness**
- Even 1 outlier can dramatically change the mean
- Mean moves towards outlier; Standard deviation is inflated
- Severely impact Least Squares Regression (the most utilized regression)

Effects of Outliers: Examples

<u>SampDate</u>	<u>PST</u>	<u>TOC</u>	
10/25/2004	1442	10.04	
10/25/2004	1612	9.90	
10/25/2004	1624	3.26	
10/25/2004	1630	3.29	Mean: 5.51
10/25/2004	1636	3.29	Median: 3.29
10/25/2004	1642	3.28	Std: 3.46

<u>SampDate</u>	<u>PST</u>	<u>TOC</u>	
10/25/2004	1624	3.26	Mean: 3.28
10/25/2004	1630	3.29	Median: 3.29
10/25/2004	1636	3.29	Std: 0.014
10/25/2004	1642	3.28	

Median is resistant to outliers. It is said to be a Robust measurement



Detecting Outliers

Several methods based on data type (univariate, multivariate, time series, etc.)

- In environmental lab data, Grubbs' method most widely used

Grubbs' statistic: $G = \frac{\max |X - \bar{X}|}{s}$ **Con:** Assumes normal distribution. Uses mean & std

- Graphical methods: scatter plots, boxplots, convex hull
- Control charts i.e. outlier if point is $>3\sigma$ from mean. **Con:** uses mean & std
- Distance-methods: Nearest-neighbor; Density based; Clustering based

Dealing with Outliers

- Remove (but do not delete) them before analysis. This assumes that you can reliably identify them
- Accommodate them by using methods resistant to outliers i.e. **robust methods**
- Robust methods generally utilize the median
- *Example*: Least Median of Squares Regression instead of Least Squares Regression (Rousseeuw and Leroy 2003)
- **Con**: Robust methods are not widely available in *popular* software

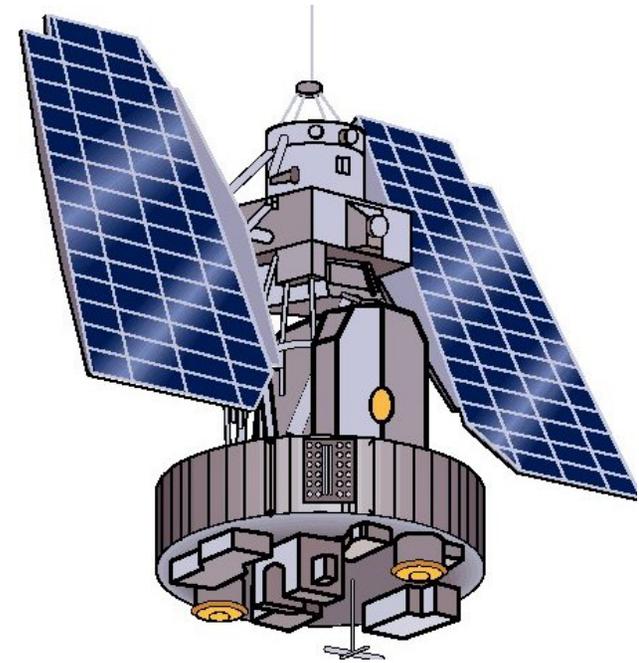
Outliers: The Good...

Outliers that are *valid data* may contain the most useful information!

- Time series: Credit card fraud detection, network intrusion (e.g. distributed denial of service) detection, fault detection
- Disease detection: x-rays, mammograms
- Data mining
- Many discoveries in science

Outliers: Ozone Hole Discovery

- In 1985, Farman et al. noticed that ozone data collected by British Antarctic Survey had fallen 10% below normal
- They were puzzled why ozone data collected by the Nimbus 7 satellite did not show this decline
- Further investigation showed that ozone concentrations recorded by the onboard Nimbus 7 instruments were so low that they were being discarded as *outliers* by the on board database programs!
- In this case, outliers were the most important data



Nimbus 7 satellite
(courtesy of NASA)

Outliers: Discovery of Argon

- In an April 1894 the British physicist Lord Rayleigh gave a lecture on generation of nitrogen
- He reported that N purified from air was slightly denser than that generated chemically due to what he postulated was a light impurity in the former
- Ramsay (a colleague) thought it was a heavier impurity in the N purified from air
- Ramsay later found out this 'impurity' (*outlier*) was a new inert gas: Argon. This led to the discovery of many other noble (inert) gases



Lord Rayleigh)

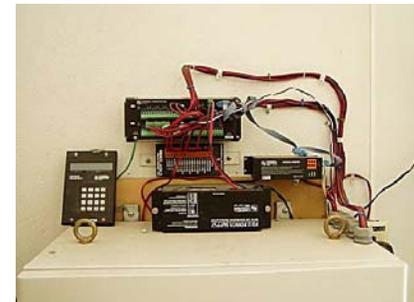
Outliers & Instrument Maintenance

Keeping record of outliers due to instrument measurement errors will assist in:

- Data validation
- Future troubleshooting
- Knowledge of system weak links (parts that break more often)
- Future instrument purchases (don't buy replacement instrument with problem history)



Water delivery/filtration system



Data logger

Literature cited

Barnett V, Lewis T. 1998. Outliers in Statistical Data. Chichester, England: John Wiley and Sons.

Grubbs FE. 1969. Procedures for Detecting Outlying Observations in Samples. *Technometrics* 11: 1-21.

Hodge VJ, Austin J. 2004. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review* 22: 85-126.

Rousseeuw PJ, Leroy AM. 2003. Robust Regression and Outlier Detection. Hoboken, New Jersey: John Wiley and Sons, Inc.