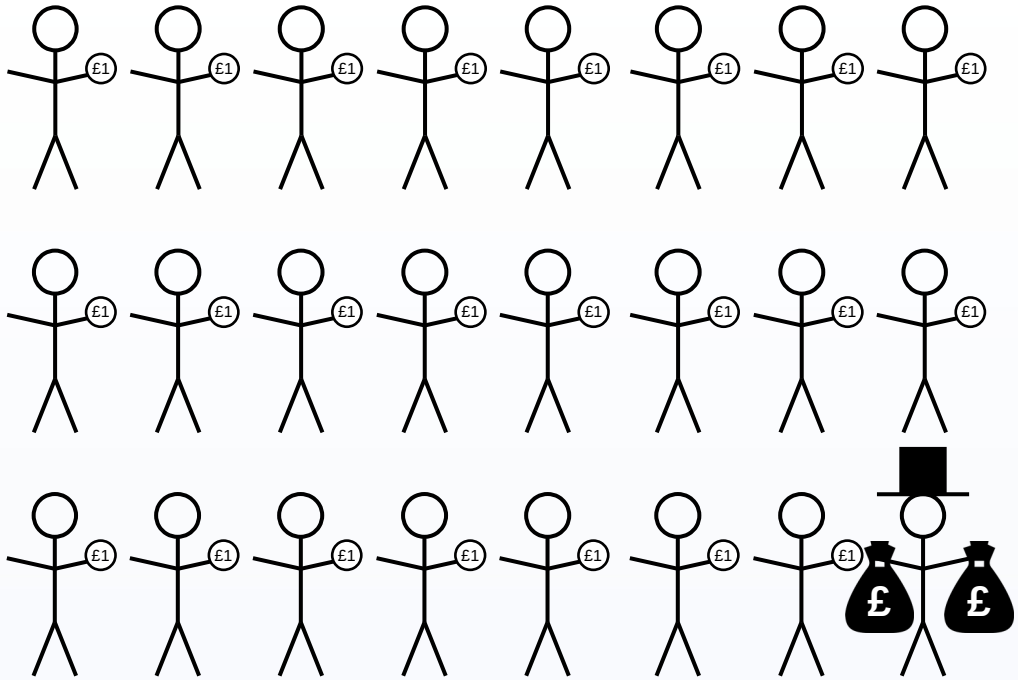
EGM101 – Skills Toolbox

Week 6, Part 5: Outliers

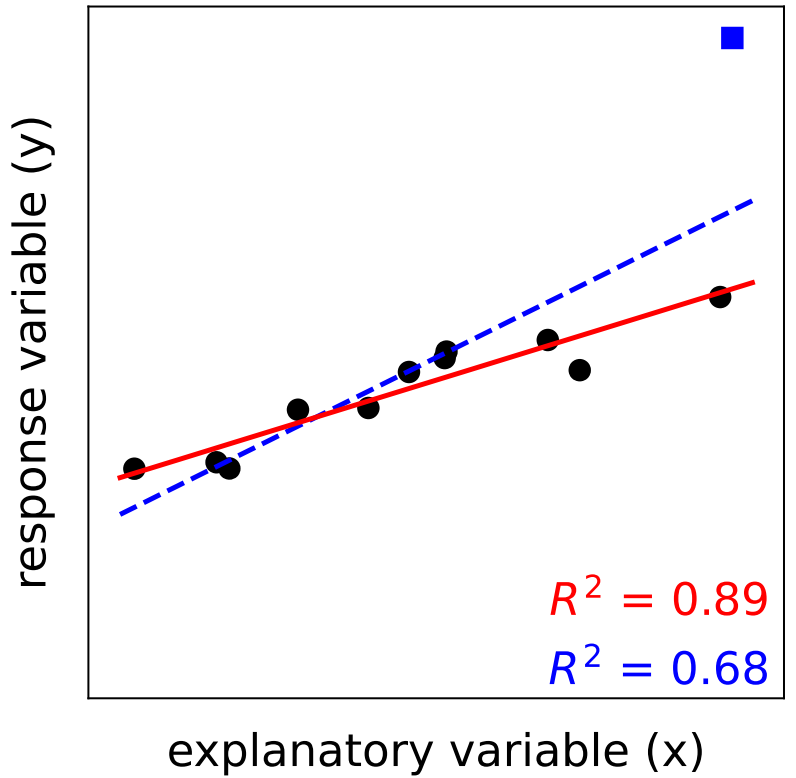
What is an outlier?

- In general: values that lie far away from the rest of the data
- Example: net worth of:
  - everyone in this class
  - <insert random billionaire>
- Here, specifically considering points that lie far away from regression line



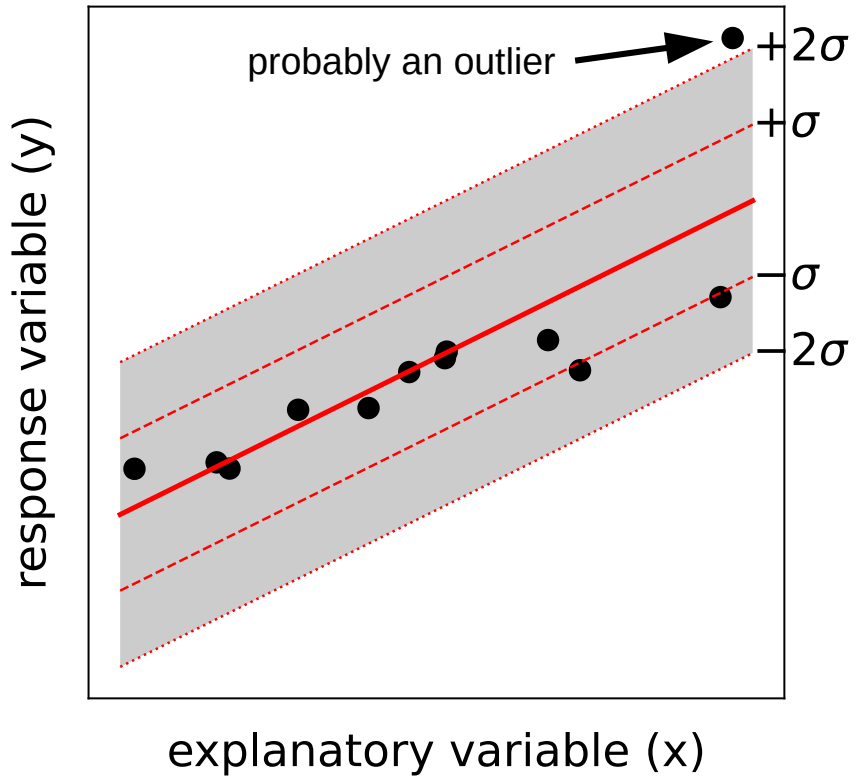
Effect of outliers on regression

- Large outliers tend to “pull” the regression line toward themselves
  - Especially if there’s a large horizontal distance (influential points)
- Remember: as distance between data, regression line increases:
  - R^2 decreases (more variability)
  - $|r|$ decreases



Identifying Outliers: Graphical

- Basic rule of thumb: points further than two standard deviations from regression line are (probably) outliers
  - Note: standard deviation of *residuals*, not the observations!
  - Note: instead of $n-1$, we divide by $n-2$ when estimating this standard deviation
- On a graph:
  - Plot data, regression line
  - Plot regression line ± 2 standard deviations
  - Points outside of gray envelope: outliers*



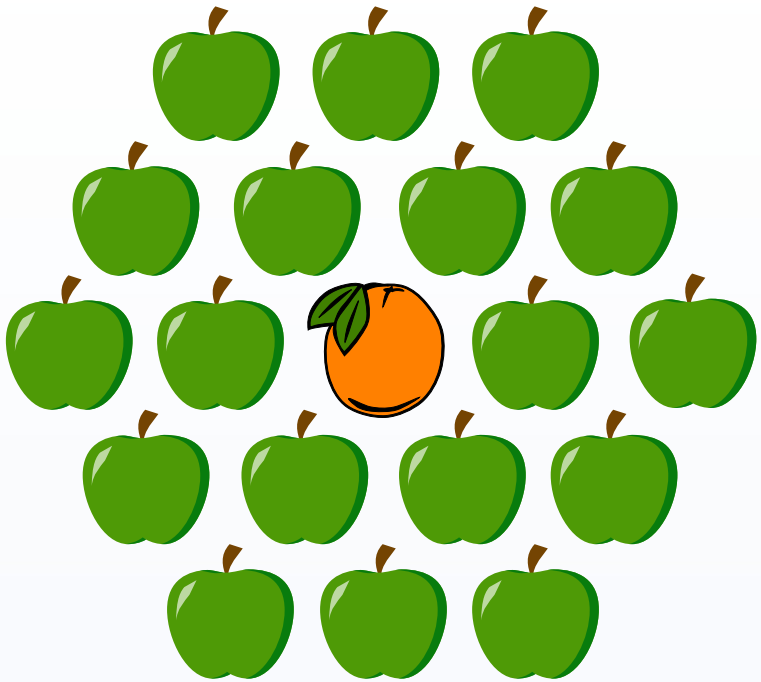
*probably.

Identifying Outliers: statistically

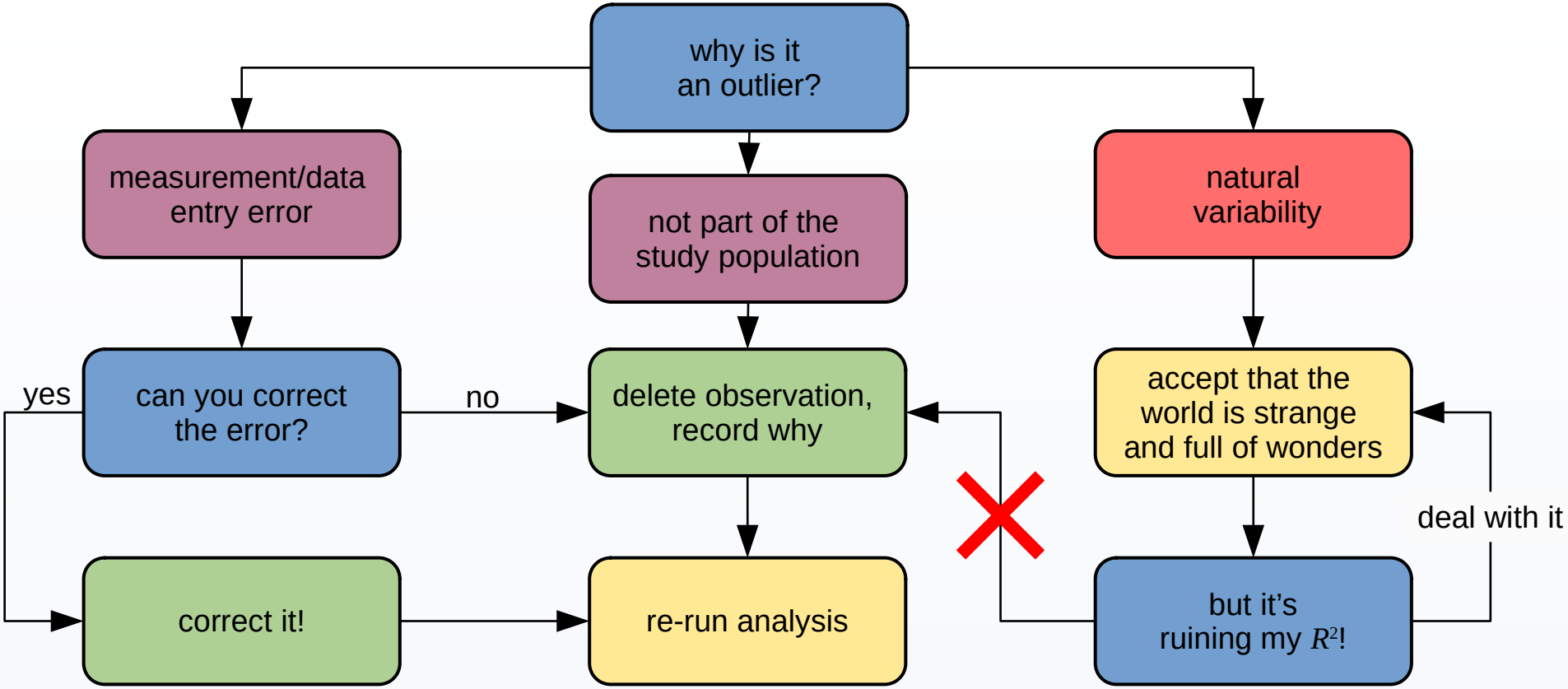
1. First, find the residuals $(y - \hat{y})$
2. Square the values you get in <1.>
3. Take the sum of <2.>, divide by $n - 2$
4. Take the square root of <3.> - this is the standard deviation.
5. Take the absolute value of the values in <1.>
6. Anything in <5.> that is larger than 2 * <3.> is probably an outlier.

Handling Outliers

- Once we've identified potential outliers, what do we do?
- First: look closer at it. Why is it an outlier?
  - Measurement/data entry error?
  - An actual difference?
    - Not part of the population
    - Just an extreme value
- What to do next depends on what makes the value an outlier.



Handling Outliers: A Flow Chart



Summary

- Outliers are observations/data that lie far away from other values
- Can identify graphically or statistically
- Once (potential) outliers are identified, need to look closer to decide what to do:
  - Correct measurement/data entry errors
  - Remove observations outside of study population
  - Accept that sometimes, things are messier than we would like.
- “It makes my model work better” is **never** a valid reason to remove an outlier.

Additional Resources

- Illowsky and Dean, Chapter 12.6
- 5 Ways to Find Outliers in Your Data [Jim Frost]
- Guidelines for Removing and Handling Outliers in Data [Jim Frost]