Outlier, Leverage And Influence

After https://newonlinecourses.science.psu.edu/stat501/node/337/

Diagnostic About Sample

- So far we have only examined diagnostics about features or columns in our data visual metaphor and their impact upon specification
- But what about rows or samples and their impact on specifications

Influential Observations

What does it mean for a data point to have high leverage?

A data point has high leverage if the value X_i that point is far away from the mean of $\{X_j\}_{j \neq i}$

What does it mean for a data point to have high influence?

A data point has high influence if removing it from the estimation while significantly change the results

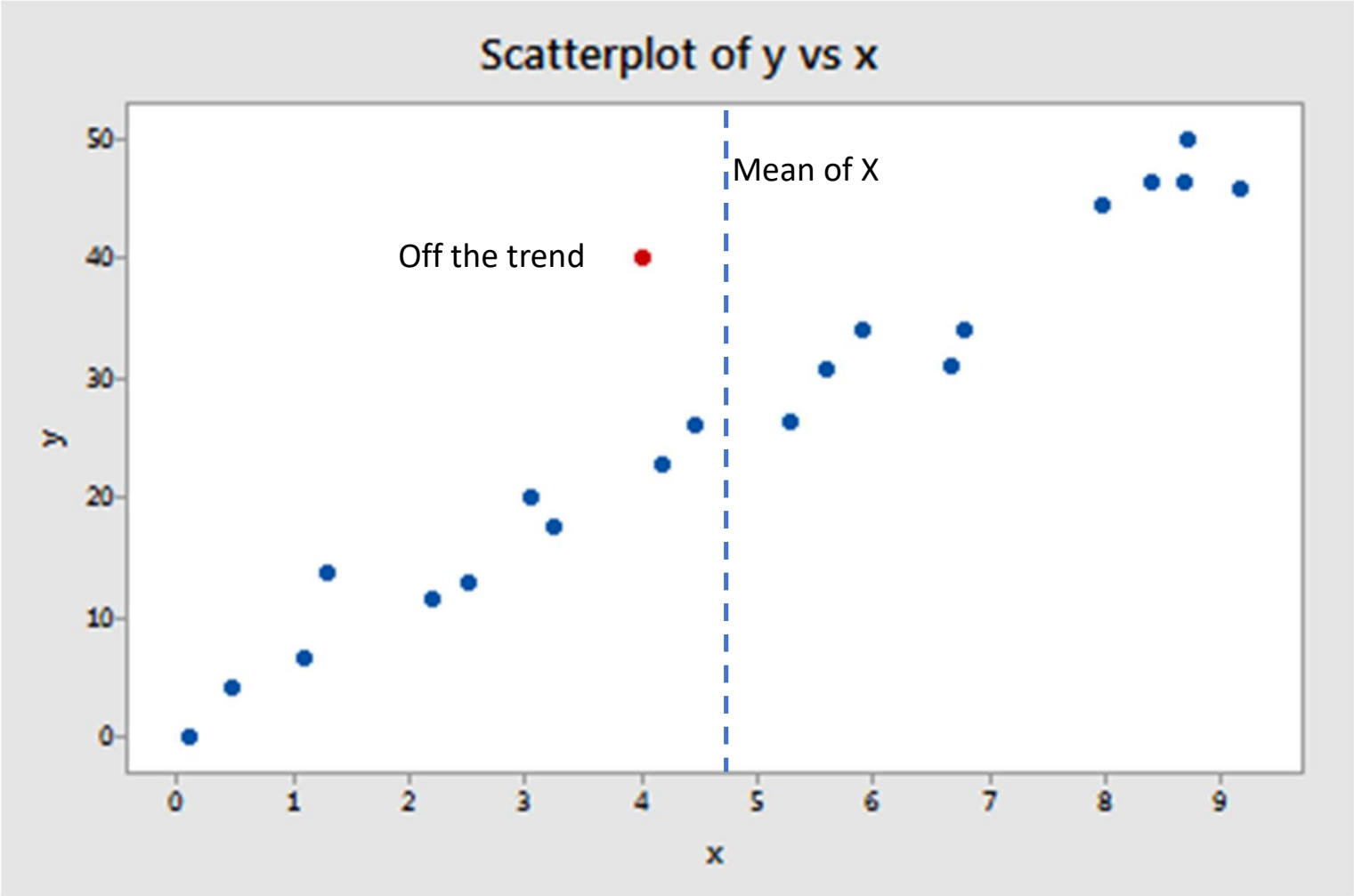
How is it possible for a data point to have low leverage but high influence?

A data point whose regressor values X are close the mean of $\{X_j\}_{j=1 \text{ to } n}$, but whose regressand Y_i value is far from the average of all other regressand values $\{Y_j\}_{j \neq i}$ (excluding itself)

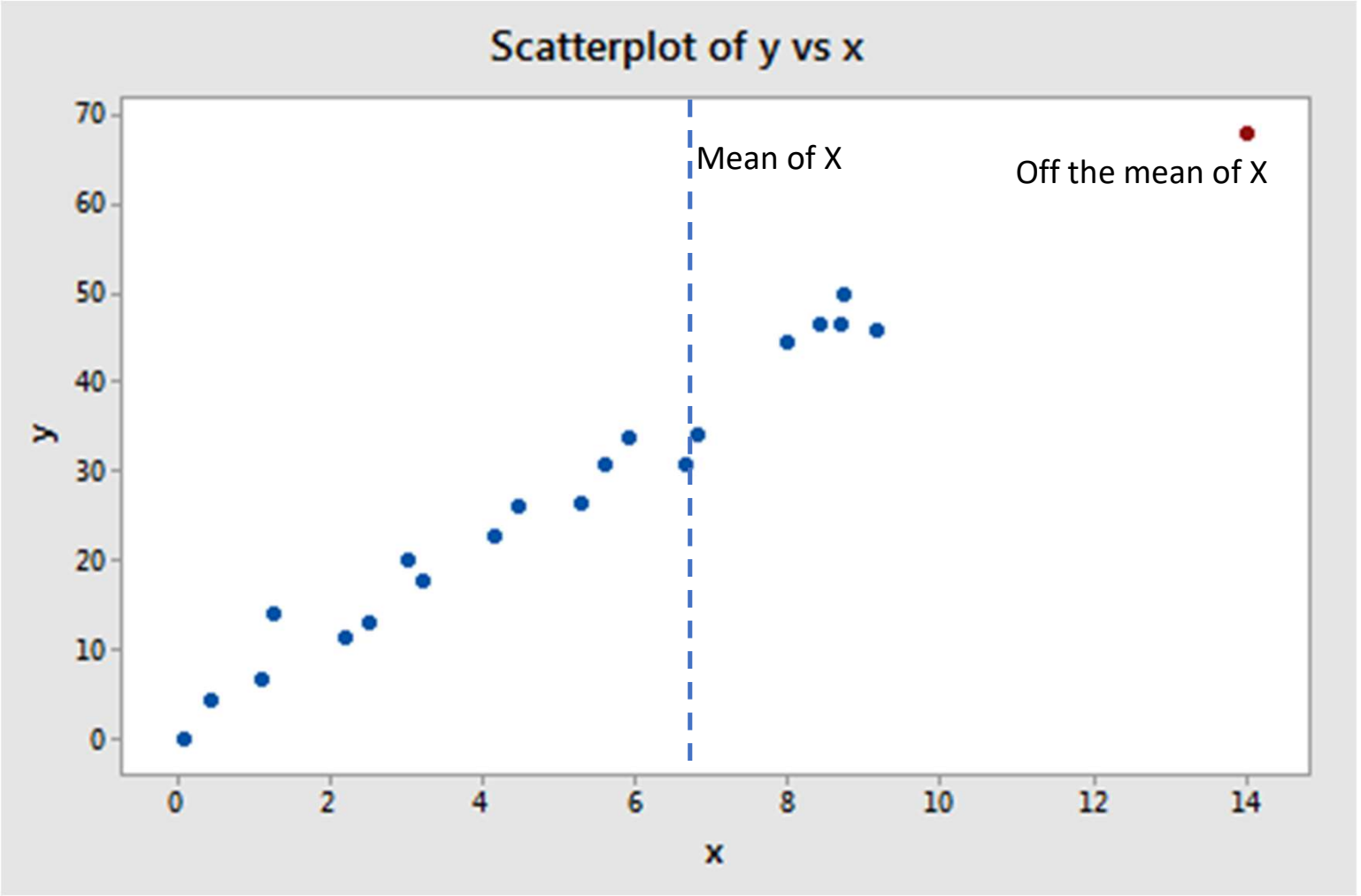
How is it possible for an outlier to have low influence?

If its regressor values X_i are far away from the mean of $\{X_j\}_{j=1 \text{ to } n}$ but the response value is close to what would be predicted if it was excluded from estimation

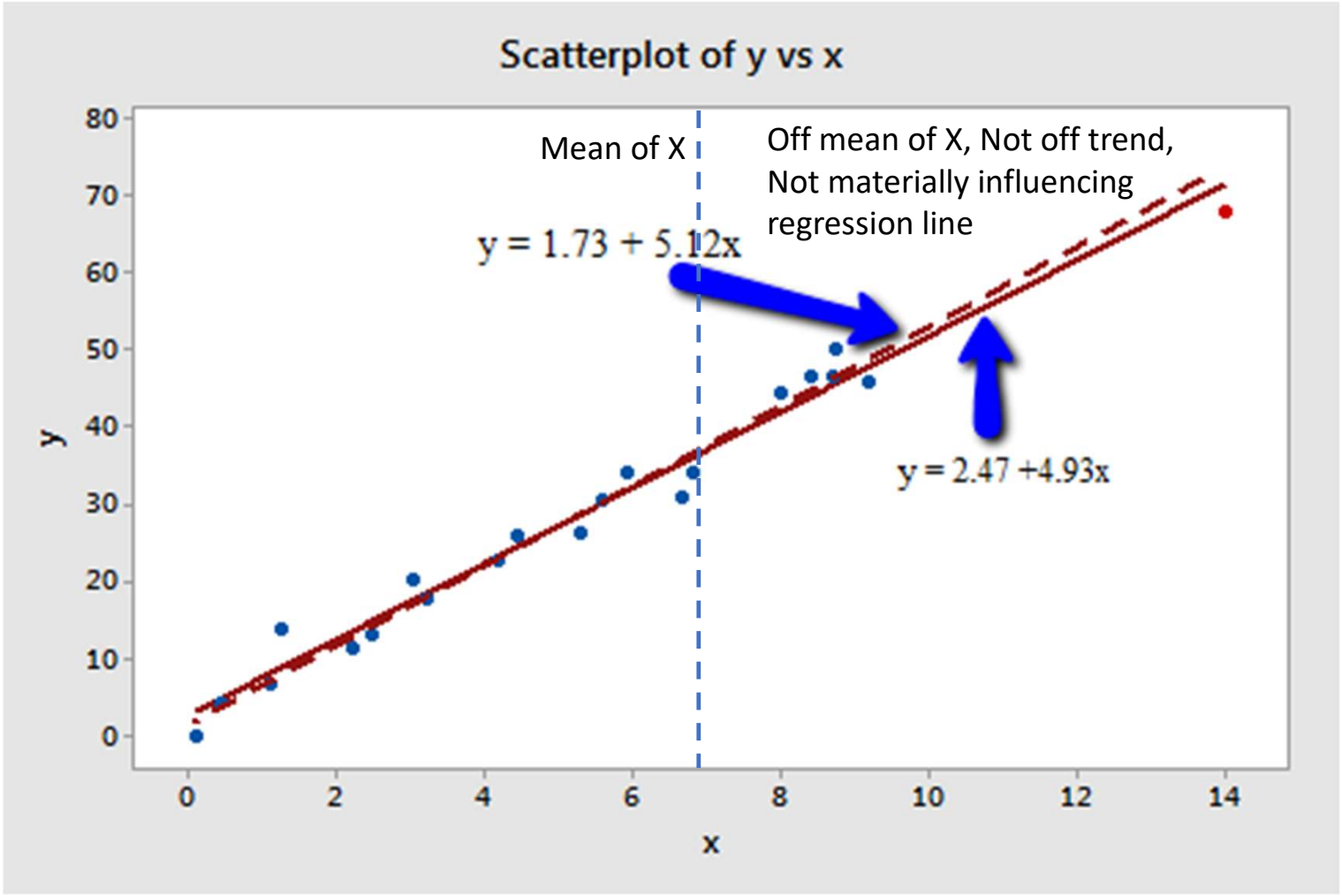
Outlier and Not Leverage



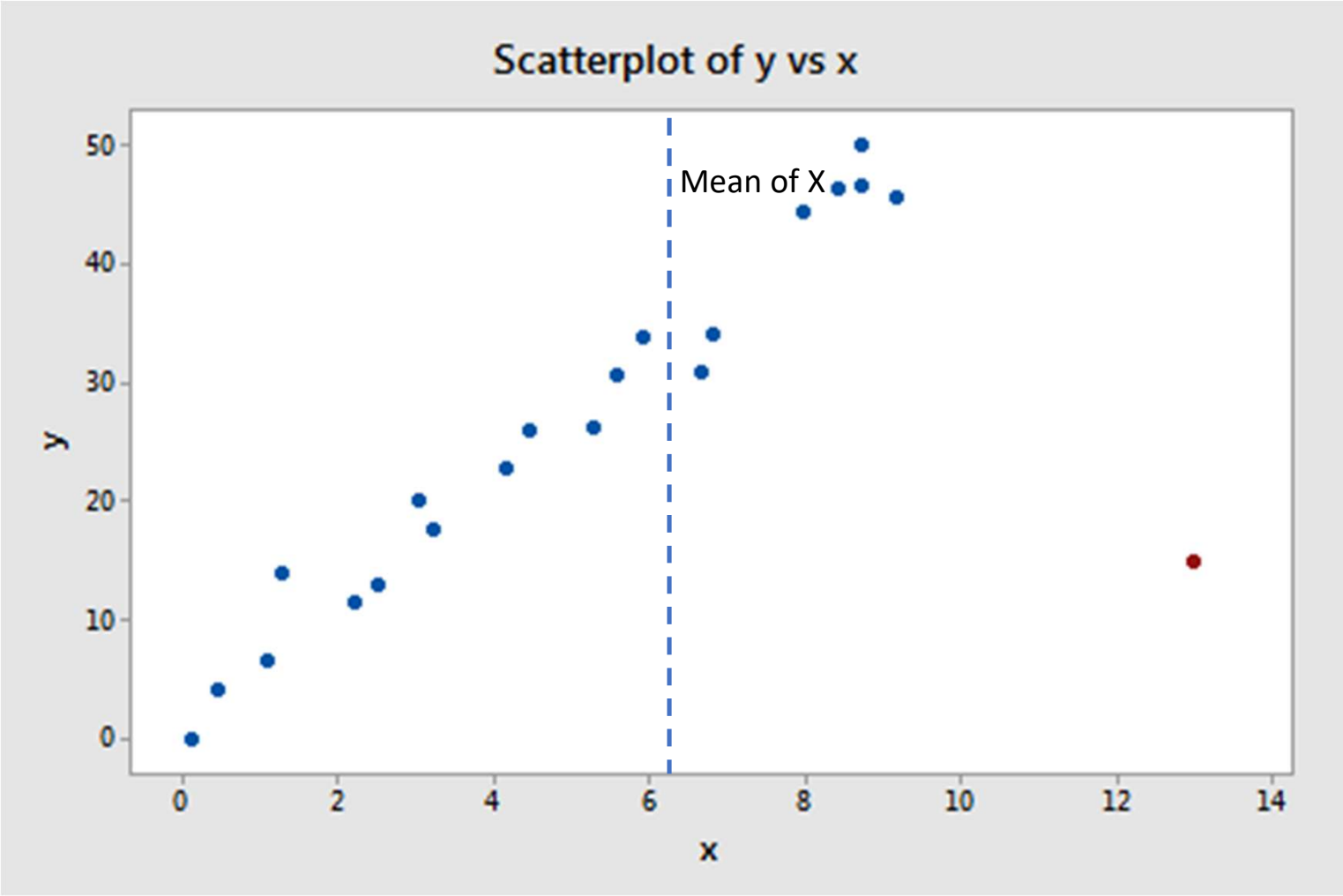
High Leverage, Not Outlier



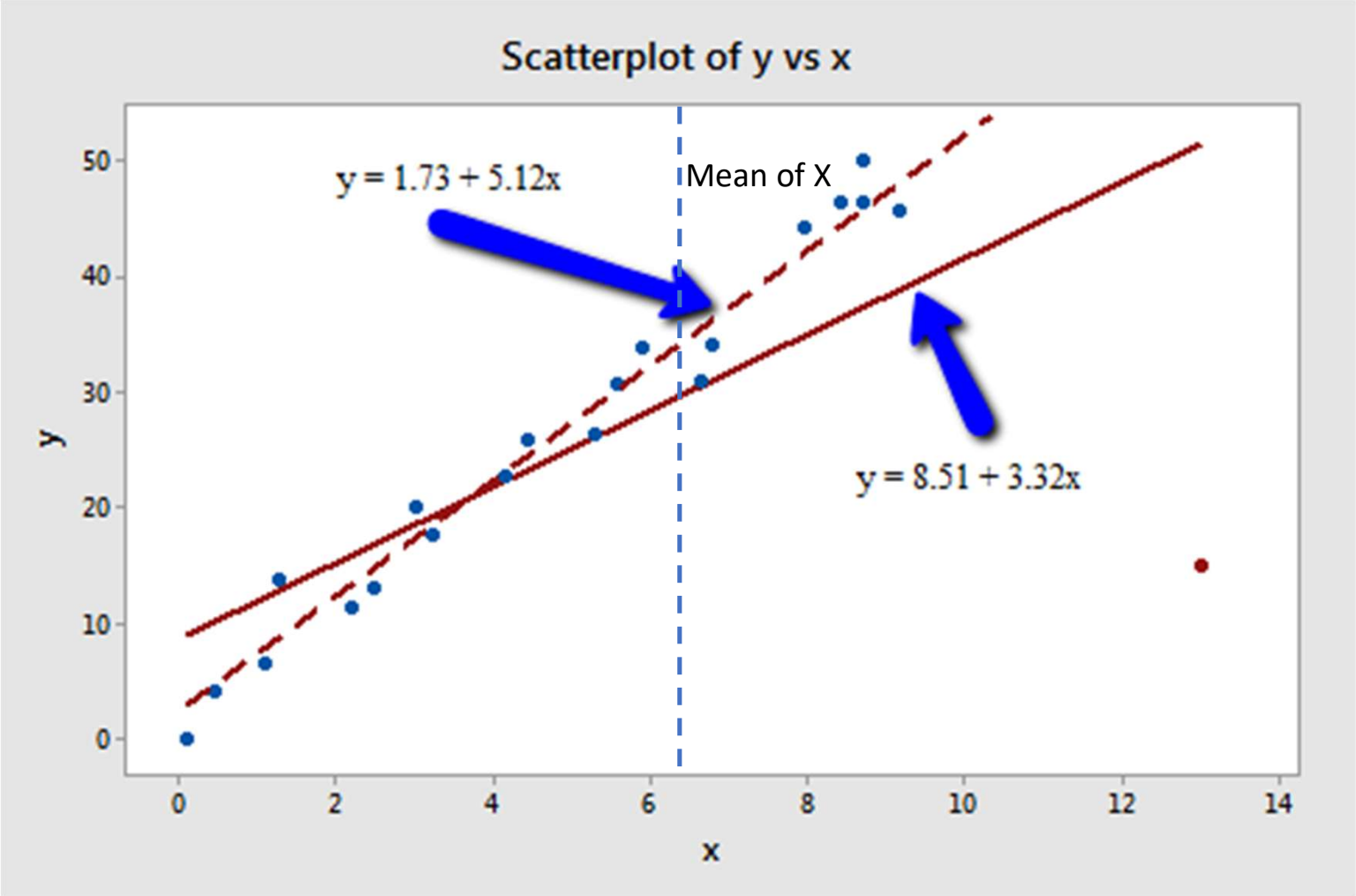
Leverage Not Outlier and Not Influence



Outlier and Leverage



Outlier Influence and Leverage



What To Do With “Outlier” Samples

- Do not throw them out without reason and examination
- Examinations
  - Transcription errors
  - Alternative population member
  - Insufficient control

- **From Wikepedia --Cook's distance** or **Cook's *D*** is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis.[1] In a practical ordinary least squares analysis, Cook's distance can be used in several ways: to indicate influential data points that are particularly worth checking for validity; or to indicate regions of the design space where it would be good to be able to obtain more data points. It is named after the American statistician R. Dennis Cook, who introduced the concept in 1977.[2][3]