Multiple Regression Diagnostics

Stat 230: Applied Regression Analysis

Warm up

With your group

- List the conditions for MLR
- Brainstorm ways to check these conditions

Regression is *not* resistant to outliers

Solid line = with outliers; dashed line = without outliers

SLR is *not* resistant to outliers

$$\hat{\beta}_1 = r \cdot \frac{s_y}{s_x} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Why?

- r is not resistant
- s_y, s_x are not resistant
- $\bar{x}, \bar{y}$ are not resistant

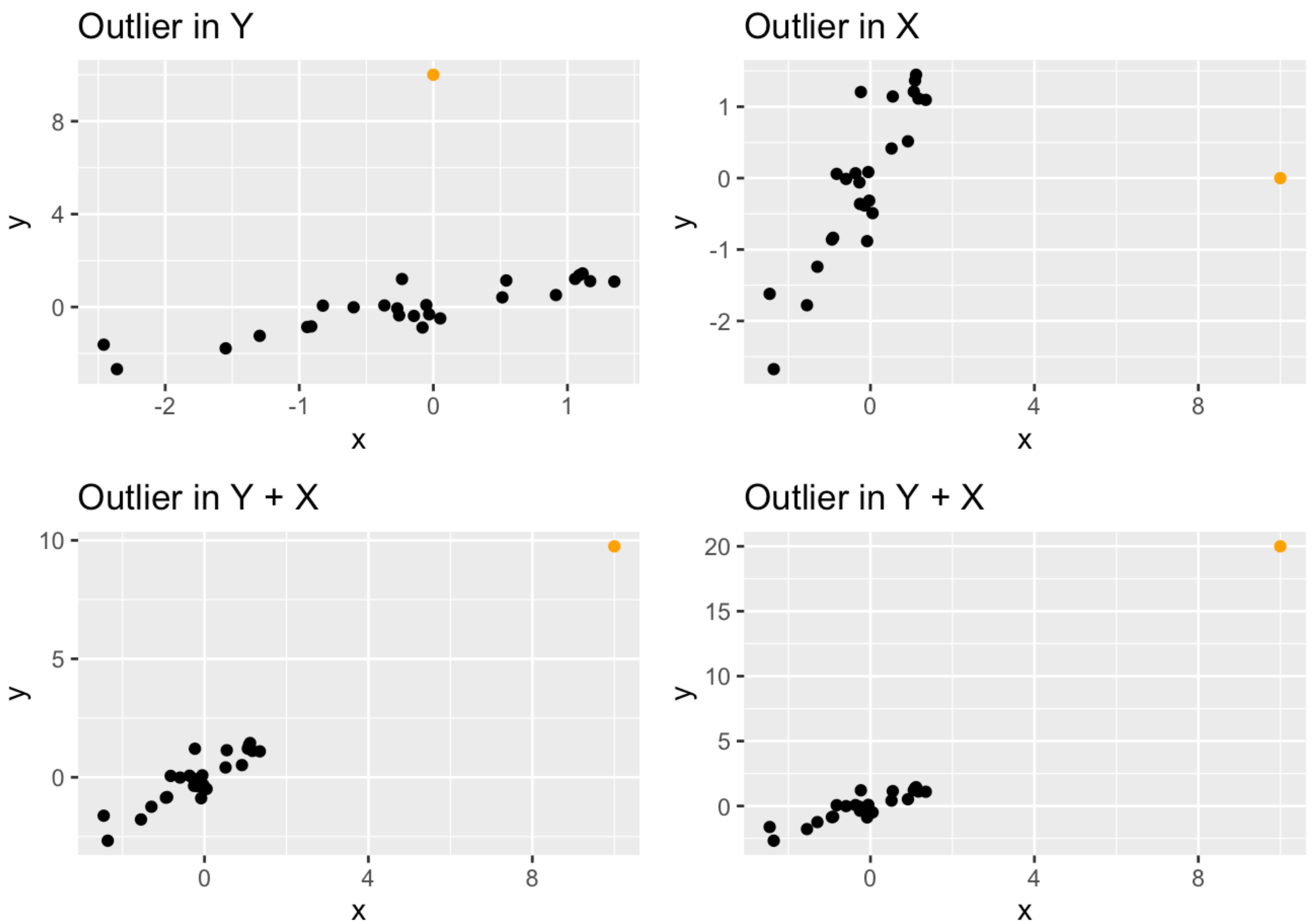
MLR is *not* resistant to outliers

We're still choosing the $\widehat{\beta}_i$ to minimize the sum of squared residuals

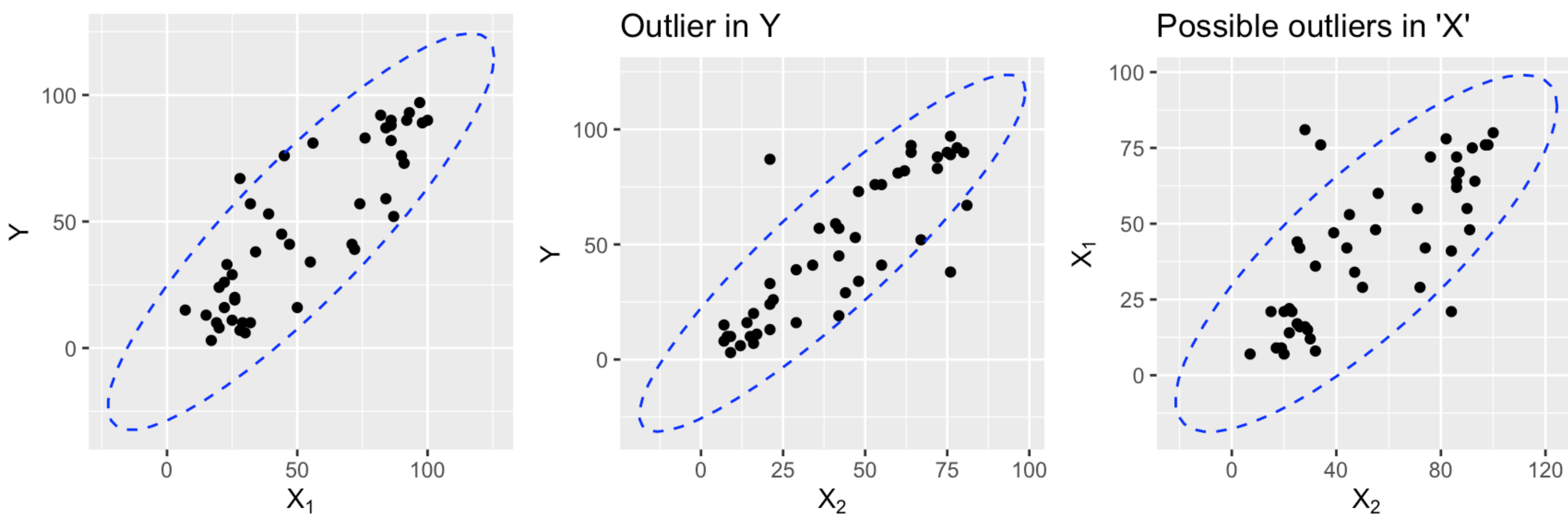
$$
\begin{aligned}
SSE &= \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2 \\
&= \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \ldots - \widehat{\beta}_p x_{ip}\right)^2
\end{aligned}
$$

Sum's aren't resistant to outliers!

Types of outliers



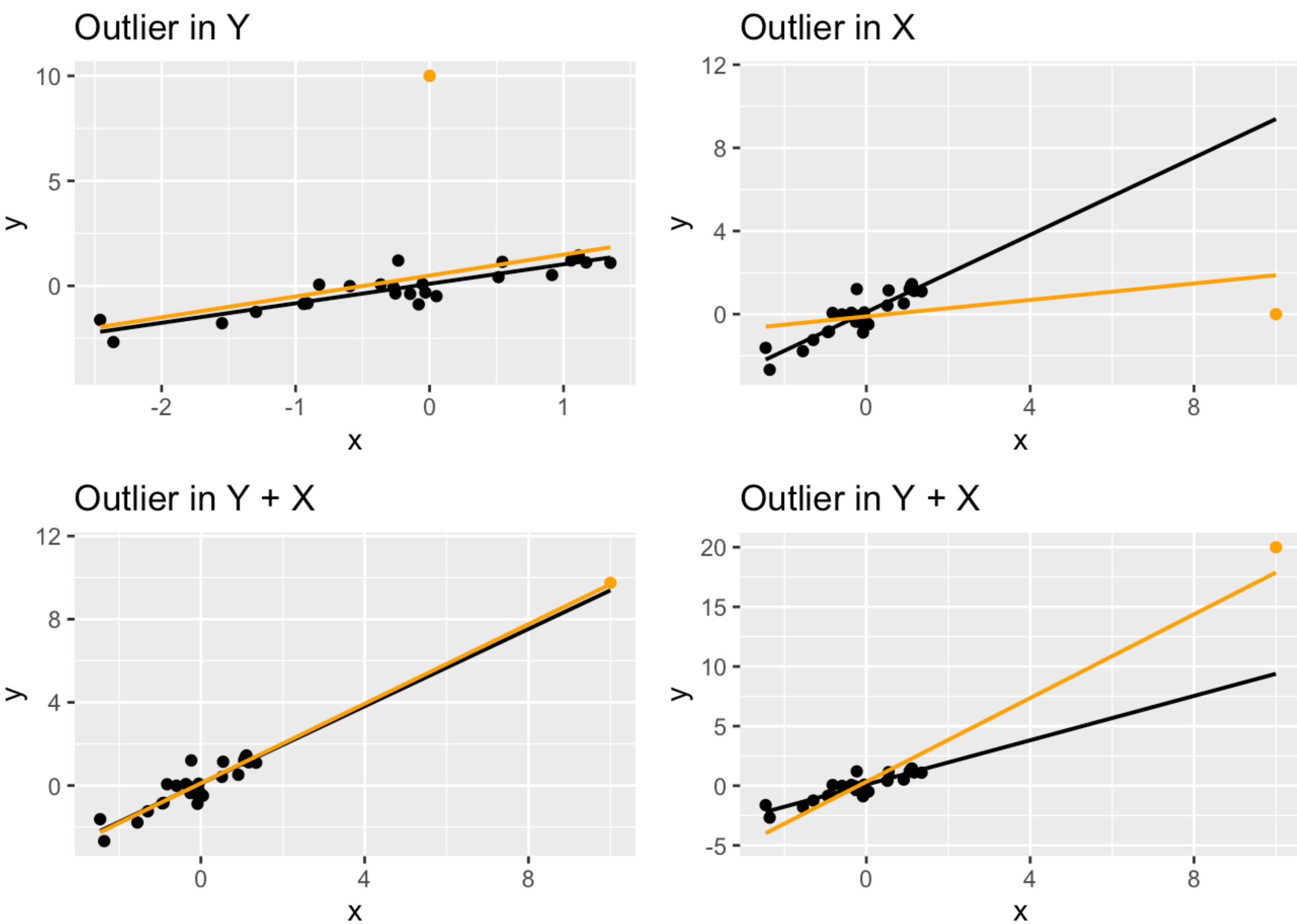
Outliers



- In higher dimensions, outliers can be tricky to detect
- To detect outliers in X, you have to consider the multivariate relationships between all of the predictors

Not all outliers are *influential*

Orange line = SLR with outlier; Black line = SLR without outlier



Influential points

Points that are able to substantially change the fitted model ares **influential**

How do we find outliers and determine if they are influential?

- Plot the data
- Plot the standardized residuals
- Fit the model with/without the point(s)
- Calculate to *influence diagnostics*

Leverage

Measures the **potential to influence** the SLR line

For SLR:

$$h_i = \frac{1}{n-1}\left[\frac{x_i - \overline{x}}{s_x}\right]^2 + \frac{1}{n} = \frac{1}{n} + \frac{\left(x_i - \overline{x}\right)^2}{\sum\left(x_i - \overline{x}\right)^2}$$

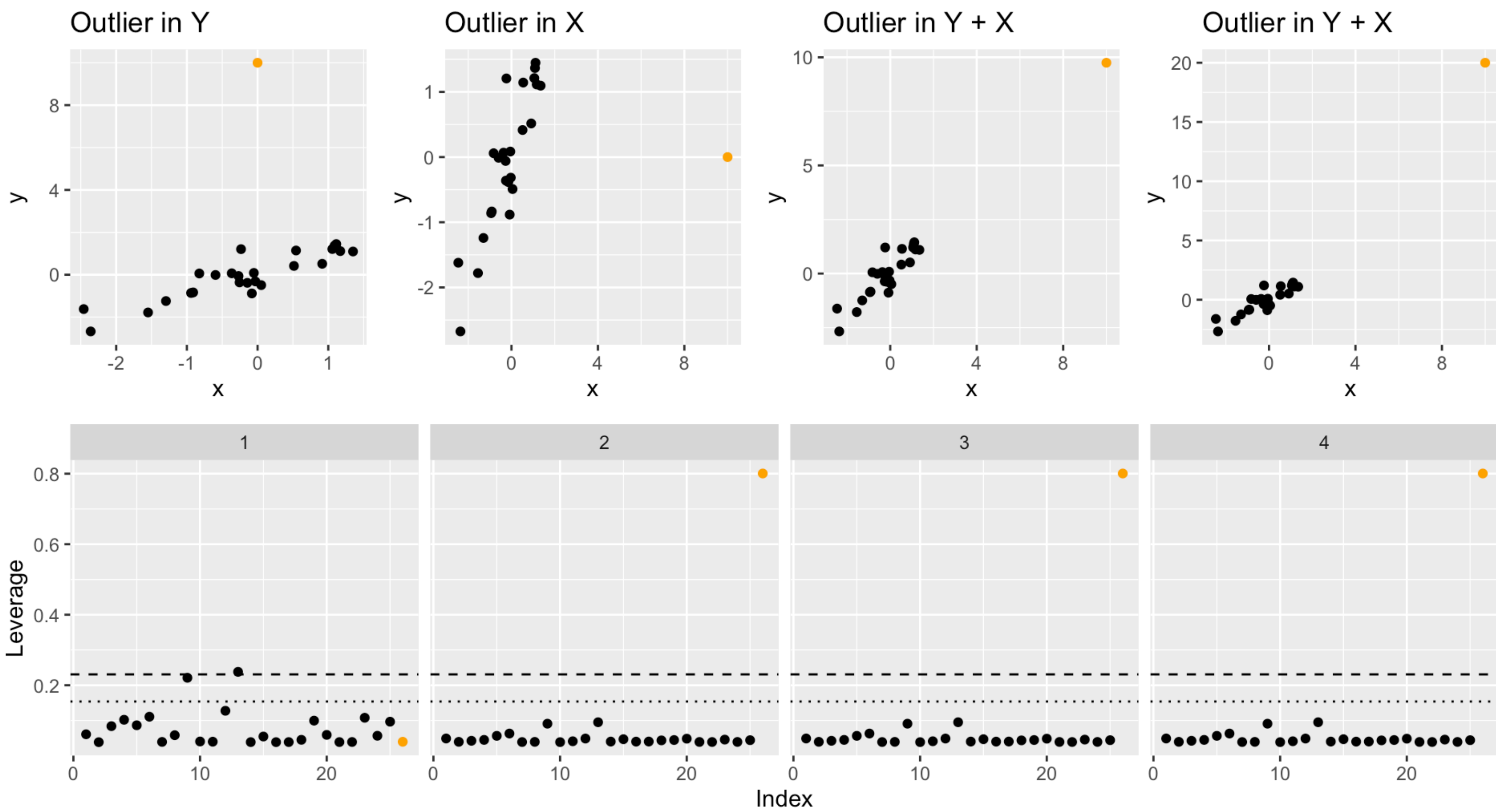
For MLR: more complicated because we're looking at a distance of a case from the average of p predictors

Cutoff

- $2(p+1)/n$
- $3(p+1)/n$
- Also a good idea to examine a histogram/index plot

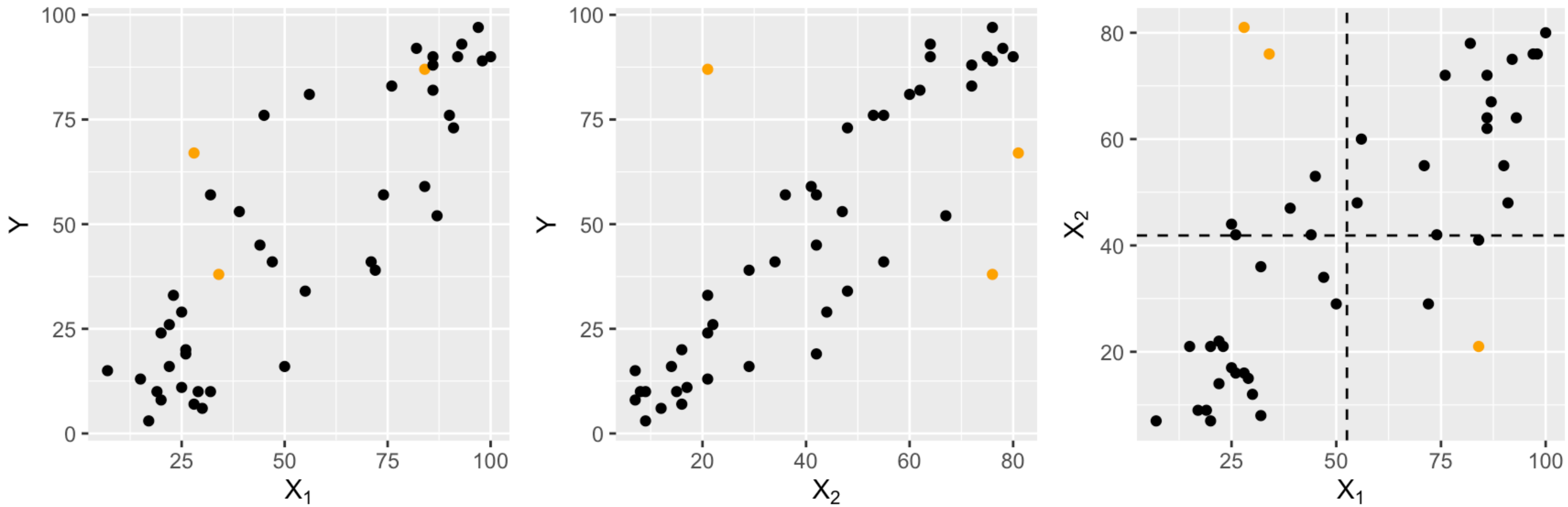
Leverage

“Flares up” when value of x is far from the mean



Leverage

In higher dimensions, leverage is focusing on the distance from the average of all of the predictors



Standardized residuals

Standardized residuals are calculated by dividing the residuals by their standard deviation

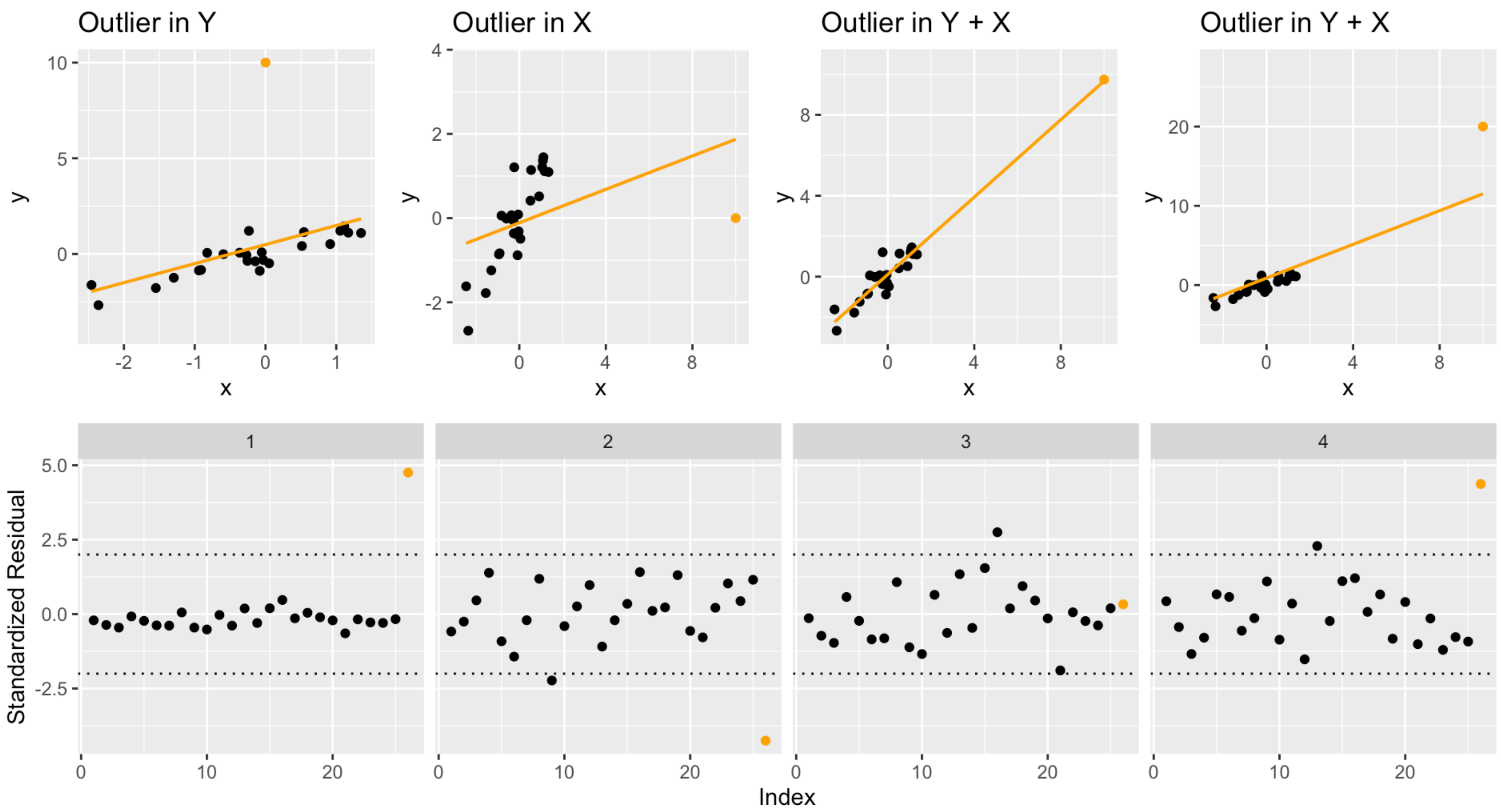
$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{\sum(x_i - \overline{x})^2}}} = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

Guidelines:

- $|r_i| > 2$ for small data sets
- $|r_i| > 4$ for large data sets

Standardized residuals

"Flare up" when value of y is far from $\hat{y}$



DFFITS

Measures effect the ith case has on its **own** fitted value

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_i}}$$

where the subscript (i) indicates that the value is based on a model when observation i is omitted

Cutoff

- 1 for small/medium data sets
- $2\sqrt{\frac{p+1}{n}}$ for large data sets

Cook's distance

Measures effect the i[th] case has on **all** of the fitted values

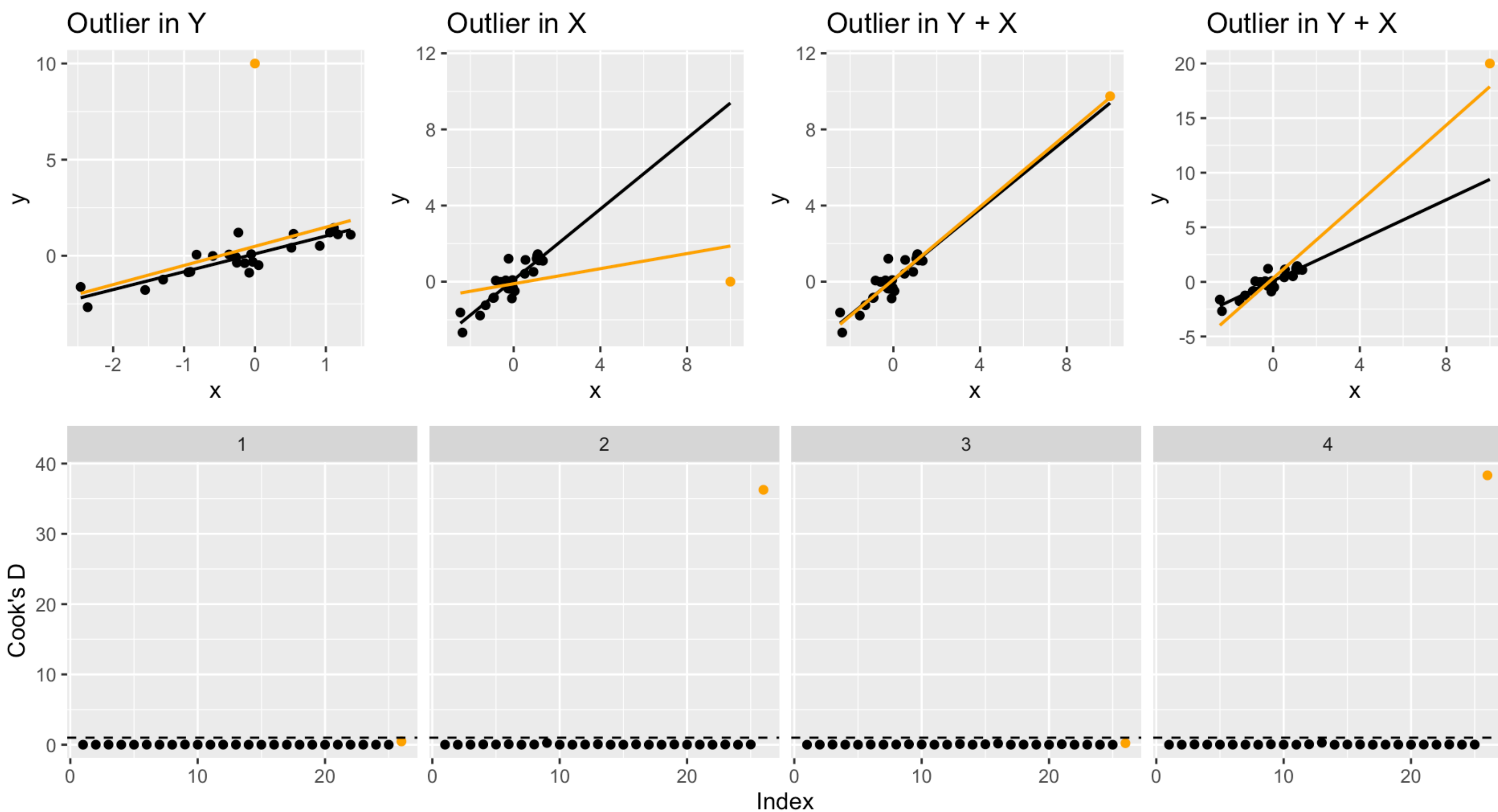
$$D_i = \sum_{j=1}^{n} \frac{\left(\hat{y}_{j(i)} - \hat{y}_j\right)^2}{(p+1)\hat{\sigma}} = \frac{r_i^2}{p+1}\left(\frac{h_i}{1-h_i}\right)$$

where $\hat{y}_{j(i)}$ is the fitted value for observation j based on a model when observation i is omitted

Cutoff

- D_i near or above 1 indicates large influence
- Find what quantile of $F_{p+1,n-p-1}$ D_i corresponds to, larger than 0.5 is cause for concern
- Also can judge relative standing of D_i — make an index plot

Cook's distance



Do we calculate these by hand?

No!

- `augment()` from the `broom` package
- `influence.measures()` from the `car` package

Advice from *Sleuth*