

Leverage and Influential points

Instructor Sara López-Pintado
Department of Statistics
Rutgers University

Leverage

- Recall that the leverage of a point is:

$$p_i = \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

- Note that the sum of the leverages are 1
- Note that we can write the formula for the slope of the regression line as:

$$b_1 = \sum_{i=1}^n p_i \frac{(y_i - \bar{y})}{(x_i - \bar{x})}$$

Leverage (2)

One of the possible lines that has slope

$$\frac{(y_i - \bar{y})}{(x_i - \bar{x})}$$

passes through (x_i, y_i) and the point (\bar{x}, \bar{y})

So, the slope of the regression line is a weighted sum of slopes of lines passing between each point and the “mean point” (\bar{x}, \bar{y})

Leverage (4)

- The weights in the weighted combination are the leverages.
- If a point has large leverage, then the slope of the regression line follows more closely the slope of the line between that point and the mean point
- This means that, for the regression, it is the points that have large leverage are important
- Points that have small leverage “do not count” in the regression – we could move them or remove them from the data and the regression line does not change very much

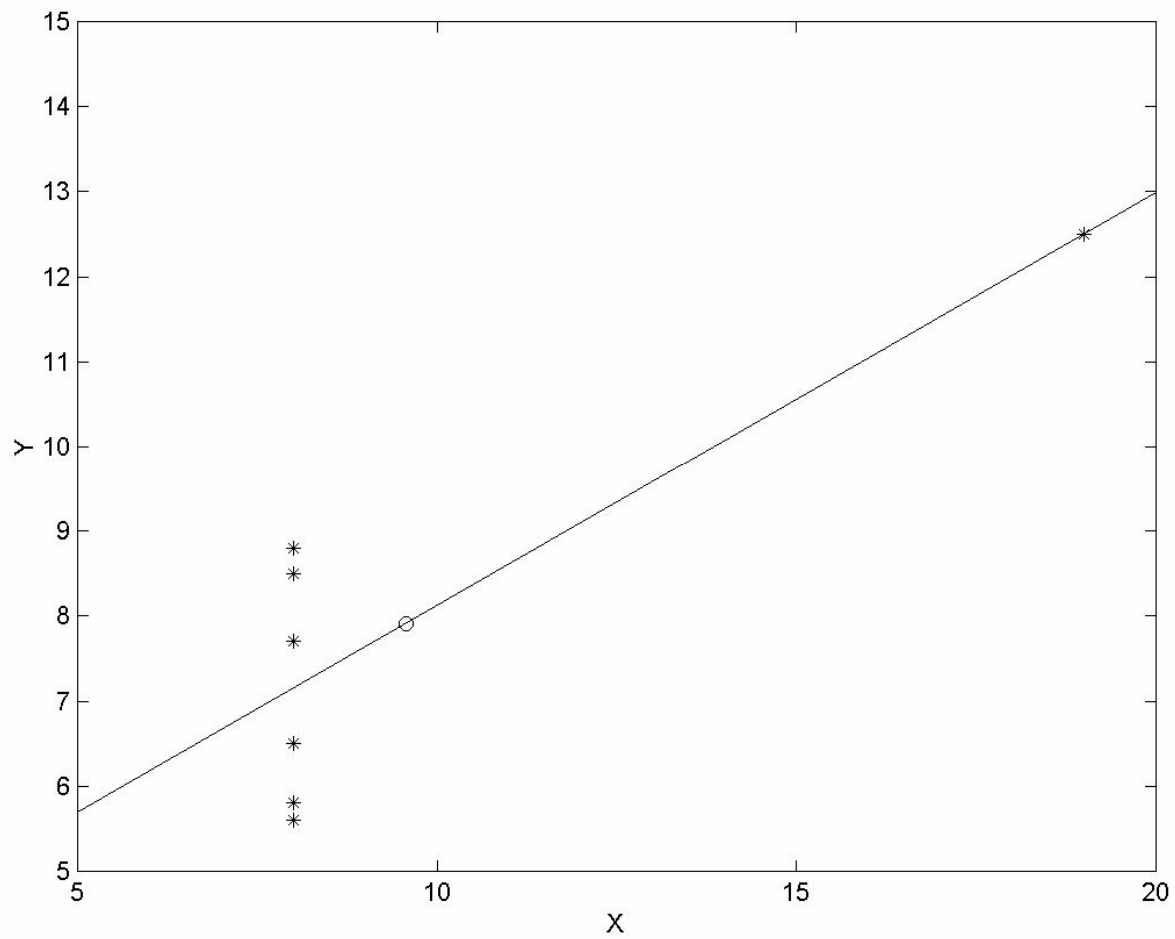
Leverage (5)

- Here is some (rather strange) regression data:

x	8	8	8	8	8	19	8
y	6.5	5.8	7.7	8.8	8.5	12.5	5.6

- The regression line is: $y = 3.26 + 0.49 x$
- On the next page, see that this line is very close to the line almost passes through (19, 12.5) and the mean point (9.6, 7.9)

Leverage (6)



Leverage (7)

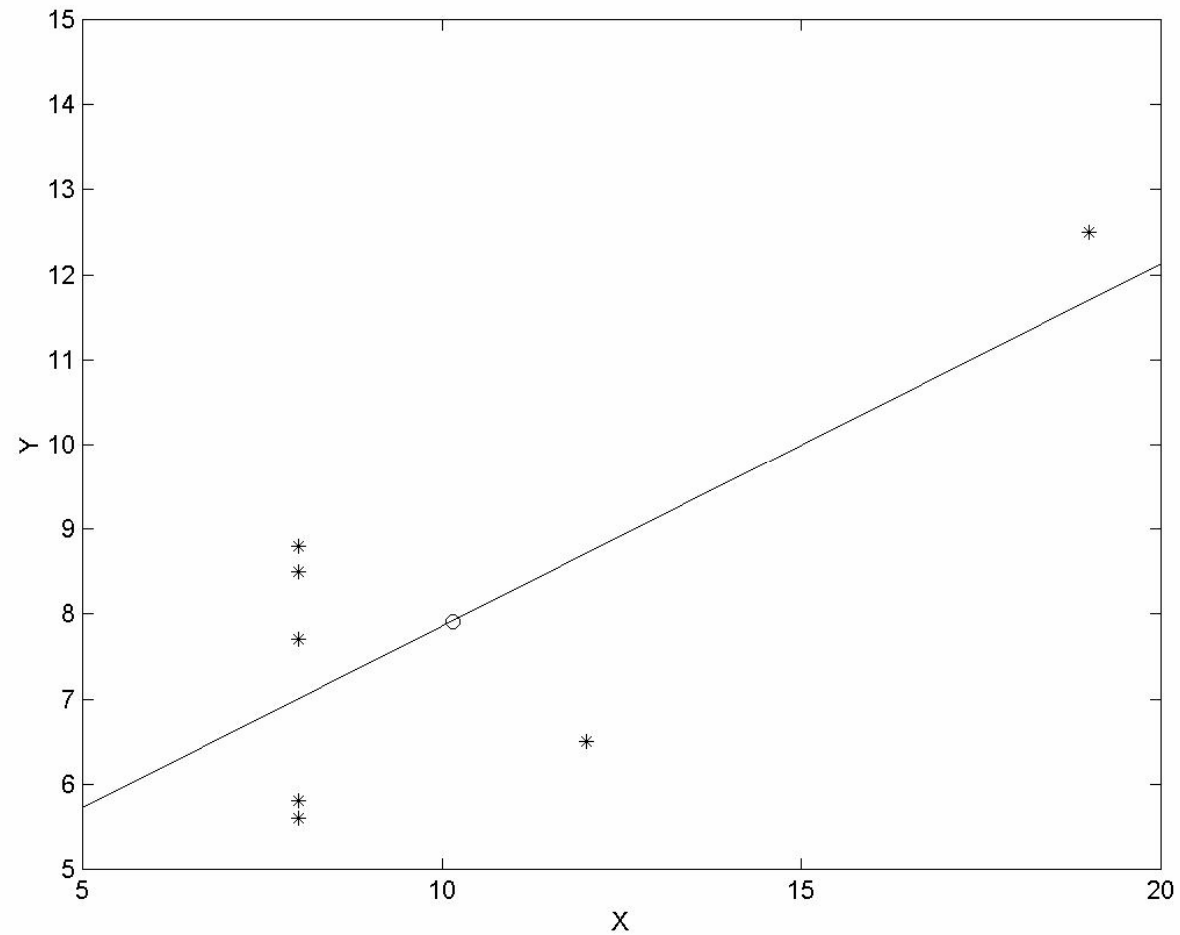
x_i	8	8	8	8	8	19	8	SUM
$(x_i - \bar{x})^2$	2.5	2.5	2.5	2.5	2.5	88.9	2.5	103.9
p_i	.024	.024	.024	.024	.024	.856	.024	1.000

$$\bar{x} = 9.571$$

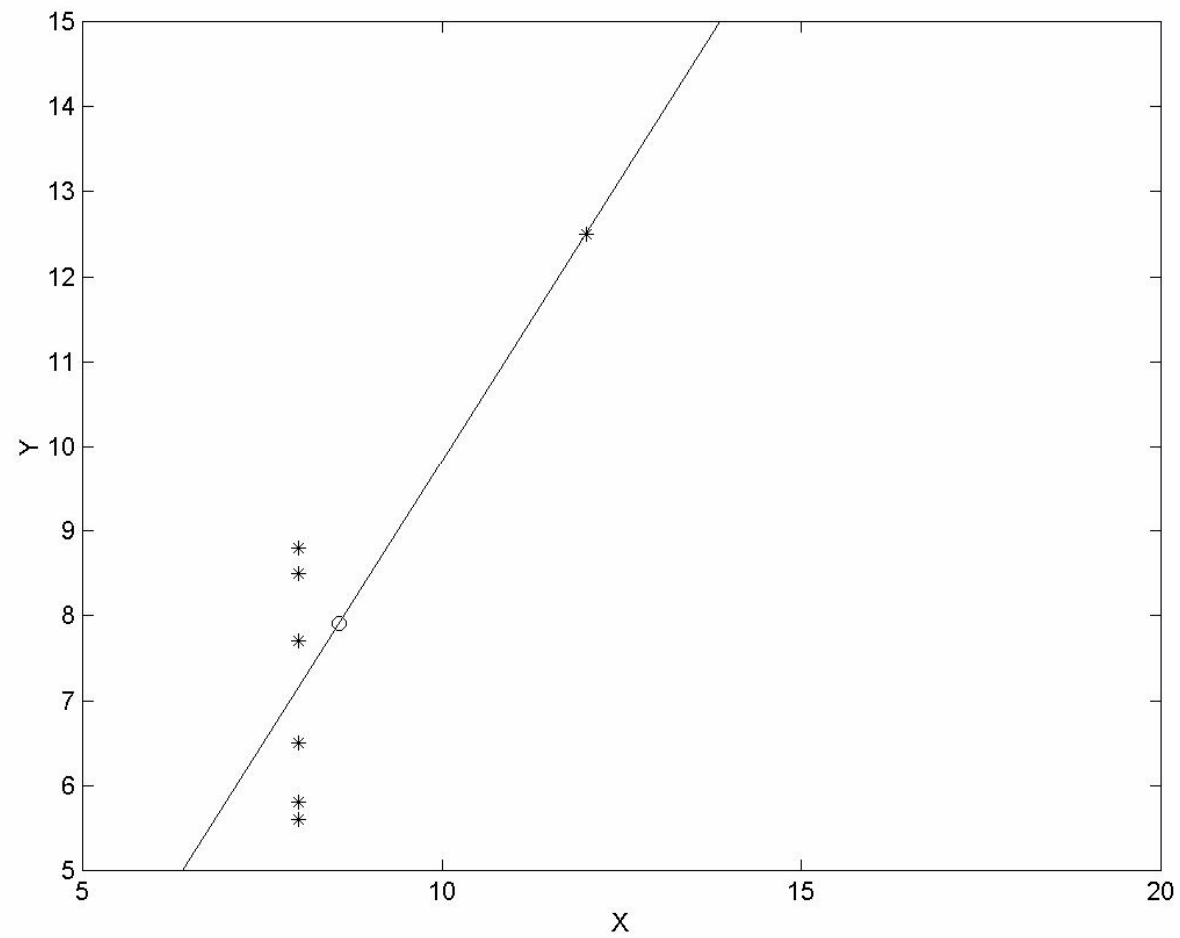
Leverage (8)

- If we move one of the points where $x = 8$, which have small leverage, the regression line changes very little
- For example, if we remove the first point $(8, 6.5)$ and fit the line to the other 6 points, we find that the regression line is $y = 3.48 + 0.47 x$
- If we move the point $(8, 6.5)$ to $(12, 6.5)$ then the regression line is $y = 3.58 + 0.43 x$
- However, if we move $(19, 12.5)$ to $(12, 12.5)$, the regression line becomes $y = -3.55 + 1.34 x$

Moving a point that has small leverage



Moving a point that has large leverage



Influential observations

- Another way to measure if an observation is important when calculating the regression line is to:
 - Calculate its predicted value
 - Remove the observation, fit the regression line without the observation, and predict its value
 - Look at the squared difference in the two predicted values, standardized by the variance in the predicted value
- If an observation is important (or influential), we see a large value in this standardized difference

Cook's distance

- Cook's distance is this measure of the influence of a point. **The Cook's distance for an observation (x_i, y_i)**

is

$$D_i = \frac{(\hat{y}_i - \hat{y}_{(i)})^2}{2 \hat{\sigma}_R^2 h_i}$$

where $y_{(i)}$ is the predicted value of y_i when the observation (x_i, y_i) is removed

- Recall that $h_i = \frac{1}{n} + p_i$
- A distance larger than 1 is an influential point

Cook's distance for the point (19,12.5)

x	12	8	8	8	8	19	8
y	6.5	5.8	7.7	8.8	8.5	12.5	5.6

- The regression line is $y = 3.58 + 0.43 x$
- The predicted value when $x = 19$ is $y_6 = 11.7$
- The residual variance estimator is **2.98**
- The leverage for this point is $p_6 = 0.86$
- Removing this point, the regression line to the other 6 points is $y = 8.84 - 0.2 x$
- The predicted value when $x = 19$ is $y_{(6)} = 5.14$

Cook's distance (2)

- Therefore the Cook's distance for (19, 12.5) is:

$$= 7.23$$

Repeating for all the other points, here are the Cook's distances:

x_i	12	8	8	8	8	19	8
y_i	6.5	5.8	7.7	8.8	8.5	12.5	5.6
D_i	<u>2.05</u>	0.08	0.03	0.17	0.12	<u>7.23</u>	0.10

Cook's distance (3)

- Points with high leverage do not necessarily have a significant Cook's distance
- Influential points are very important, since they influence the regression line very strongly.
- Influential points should be investigated. They may be outliers, at the least their value affects the regression a lot, so we must be sure that their value is correct.

Cook's distance (4)

- Usually, the presence of an influential point can mean:
 - The model is correct, but there has been an error in observing the value of the influential point
 - The value of the influential point is correct, but the model is not correct; it cannot model the influential point well
- If the model is not correct, this is usually because:
 - The relationship between x and y is not linear in an interval of values of x that includes the influential point
 - There is heteroskedasticity
 - There is another explanatory variable for y that takes a different value for the influential point

Studentized residuals (1)

- Finally, there is another way to see if an observation is an outlier.
- We already have the standardized residual:

$$\frac{e_i}{\sqrt{\hat{s}_R^2(1-h_i)}}, \text{ where } h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{ns_x^2}$$

- One problem is that if our observation is influential and/or an outlier, it increases the variance estimate a lot
- Thus the standardized residual may be smaller than we like

Studentized residuals (2)

- To solve this problem, we can estimate σ^2 without the observation.
- Let $\hat{S}_{R(i)}^2$ be the residual variance estimator from fitting the regression to $n - 1$ points without (x_i, y_i)
- The **studentized residual** is

$$t_i = \frac{e_i}{\sqrt{\hat{S}_{R(i)}^2 (1 - h_i)}}, \text{ where } h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{nS_x^2}$$