STAT 410 - Linear Regression

Lecture 10

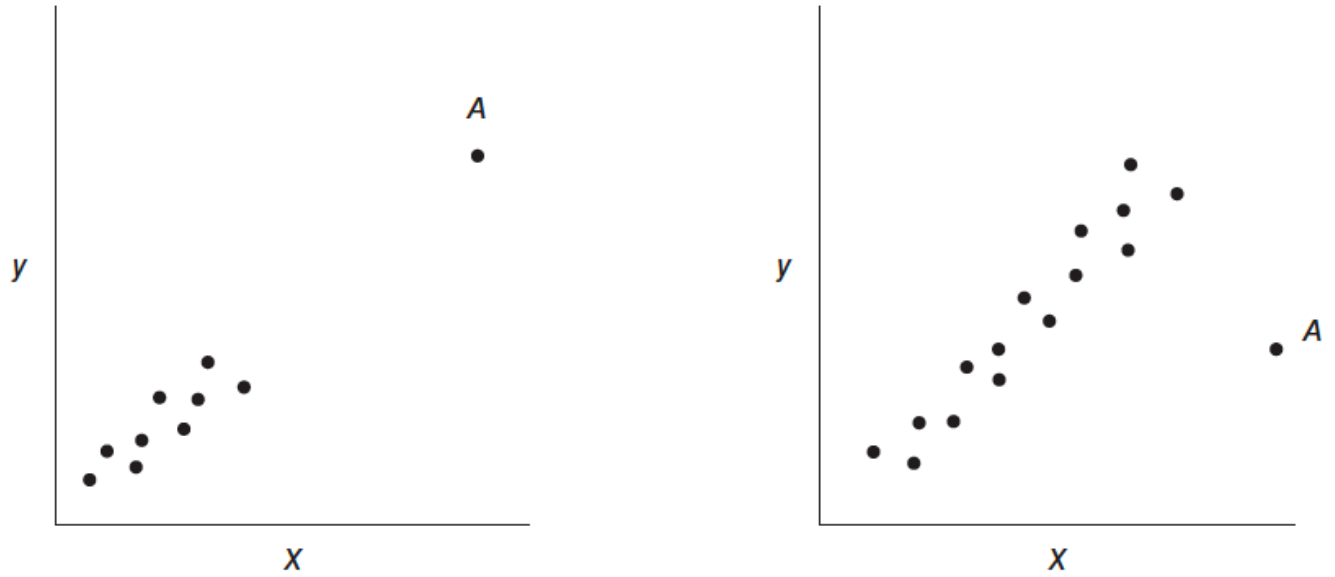
Meng Li

Department of Statistics

Oct. 17, 2017

Outliers and leverage points

- Outliers are a fact of life. We can have both x and y outliers.



- The plot on the left shows a **leverage** point. The plot on the right shows an **outlier**.
- Leverage exposes the potential role of individual data points.
- Outliers are data points which break a pattern.

Leverage

- The quantity $\frac{d\hat{y}_i}{dy_i}$ measures “leverage" of a point:
  - Let us perturb y_i a little bit at fixed x_i. How much do you expect $\hat{y}_i$ to move?
  - If $\hat{y}_i$ moves as much as y_i then clearly y_i has the potential to drive the regression - so y_i is leveraged.
  - If $\hat{y}_i$ hardly moves at all then clearly y_i has no chance of driving the regression.
- Let h_{ij} be the (i,j)th entry of $\mathbf{H}$. The fact $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ (why?) can be rewritten as

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j.$$

- Thus $h_{ii} = \frac{d\hat{y}_i}{dy_i}$ and it is defined as the **leverage** of (x_i, y_i).

Properties of leverage

- The leverage $h_{ii} = \boldsymbol{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_i$, where $\boldsymbol{x}_i$ is the ith row of $\mathbf{X}$.
- Leverage of a point only depends on x. It is a standardized measure of the distance of x_i from the center of x space.
- For example, we can obtain the following result in SLR:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}.$$

- Properties of h_{ii}:
  - $0 \leq h_{ii} \leq 1$
  - Fixed i: $\sum_{j=1}^{n} h_{ij} = 1$
  - $\bar{h} = \sum_{i=1}^{n} h_{ii}/n = p/n$
  - Fixed i: $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$
- It follows that $h_{ii} \approx 1 \Longrightarrow h_{ij} \approx 0 \Longrightarrow \hat{y}_i = y_i$. Thus if $h_{ii} \approx 1$, the fitted regression line (almost) goes through (x_i, y_i).
- Traditionally, a **leverage point** is one satisfies that

$$h_{ii} \geq 2\bar{h} = \frac{2p}{n}.$$

Outlier detection

- We have seen how we can use residuals for identifying problems with normality, constancy of variance, linearity.
- Standardized residuals, allow the residuals to be compared on the “standard scale”.
- However, the standardized residuals $s_i = y_i - \hat{y}_i$ will be influenced if y_i is really leveraged as it will drag the regression line toward it.
- Solution: use the idea of “leave-one-out" (or “jackknifing") and calculate $e_{(i)} = y_i - \hat{y}_{(i)}$, where $\hat{y}_{(i)}$ is the fitted value at x_i excluding y_i.
- It turns out

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}.$$

Studentized residuals

- We standardize $e_{(i)}$ by its variance $\sigma^2/(1-h_{ii})$: $\frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$
- This leads to the (internally) **studentized residuals**

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}}.$$

- The **externally** studentized residuals use the MS_{Res} excluding y_i.
- On the other hand, r_i is nothing but $e_i/se(e_i)$ because

$$Cov(\mathbf{e}) = \sigma^2(\mathbf{I}-\mathbf{H}) \Longrightarrow Var(e_i) = \sigma^2(1-h_{ii}).$$

- We shall use $|r_i|$ to help detect outliers.
- Many other tools to detect influence of a point: Cook's D (deletion diagnostic), DFFITS, DFBETAS, etc.