ST 517: Data Analytics I

Case Influence Measures

Outline

Case Influence Measures

- Leverage
- Studentized Residuals
- Cook's Distance

Case Influence Statistics

Case influence statistics can be used to determine whether there are certain points in our dataset that are particularly unusual and might be heavily influencing the results of the analysis.

There are *many* case influence measures: we will talk about three measures in particular.

- Leverage
- Cook's distance
- Studentized Residuals

In each case, we'll talk about what the measure attempts to identify, and how we might use it to diagnose unusual observations.

Leverage

Leverage: The **leverage** of a point (observation) is a measure of how far the values of the predictor variables for that point are from the cloud of typical predictor variable values.

A point has *high leverage* if it has a very unusual combination of predictor variable values.

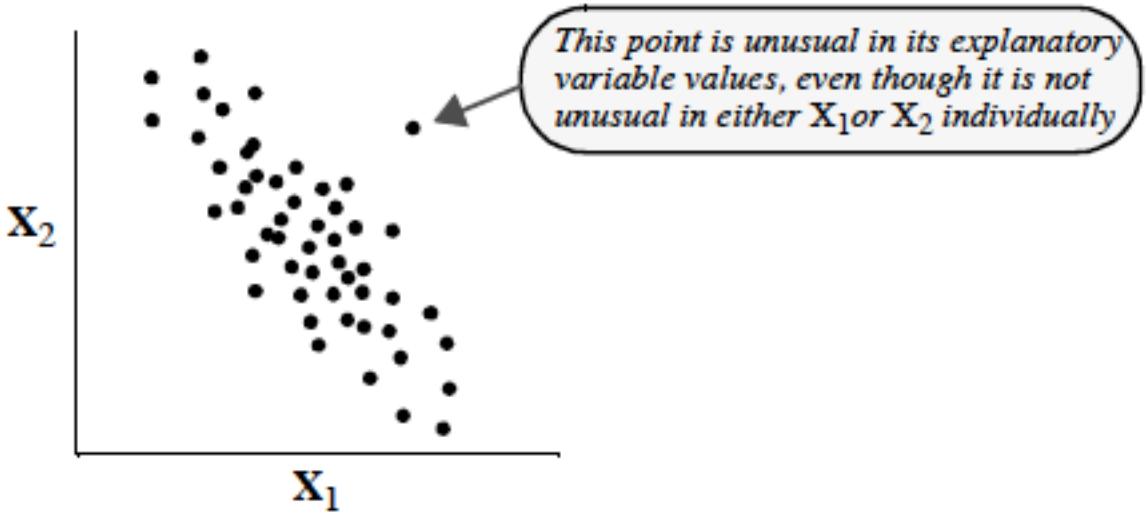
The leverage of point *i* is often denoted by h_i (and is sometimes referred to as the *i*th diagonal of the *hat matrix*).

It can be shown that $\frac{1}{n} \leq h_i \leq 1$, so values of h_i that are closer to 1 are considered large.

Leverage

Display 11.10 **p. 316**

An illustration of what is meant by "far from the average" of multiple explanatory variables when they are correlated



Understanding Leverage

Cases with high leverage have more pull on the regression fit, which means that the *residual* for a case with high leverage must be small.

Let $\hat{Y}_{(i)}$ be the fitted value for observation *i* when the fit is estimated from all of the other observations *excluding* observation *i*.

- High leverage (h_i large) and Y_i close to $\hat{Y}_{(i)}$: Observation *i* does not *influence* (change) the fit very much (since the fit already agrees with this observation).
- High leverage (h_i large) and Y_i far from $\hat{Y}_{(i)}$: Observation *i does influence* (change) the fit (the fit is pulled to agree with this observation more).

Studentized Residuals

Studentized Residuals: The **studentized residual** of a point is a rescaling of the residual to put them on a common scale.

Points that have higher leverage tend to have lower residuals, so we want to account for this tendency by rescaling (increasing) the residuals for such observations.

Points with a high studentized residual are observations that aren't well fit by the regression model. There is something unusual about their combination of explanatory variable values and their response value.

Cook's Distance

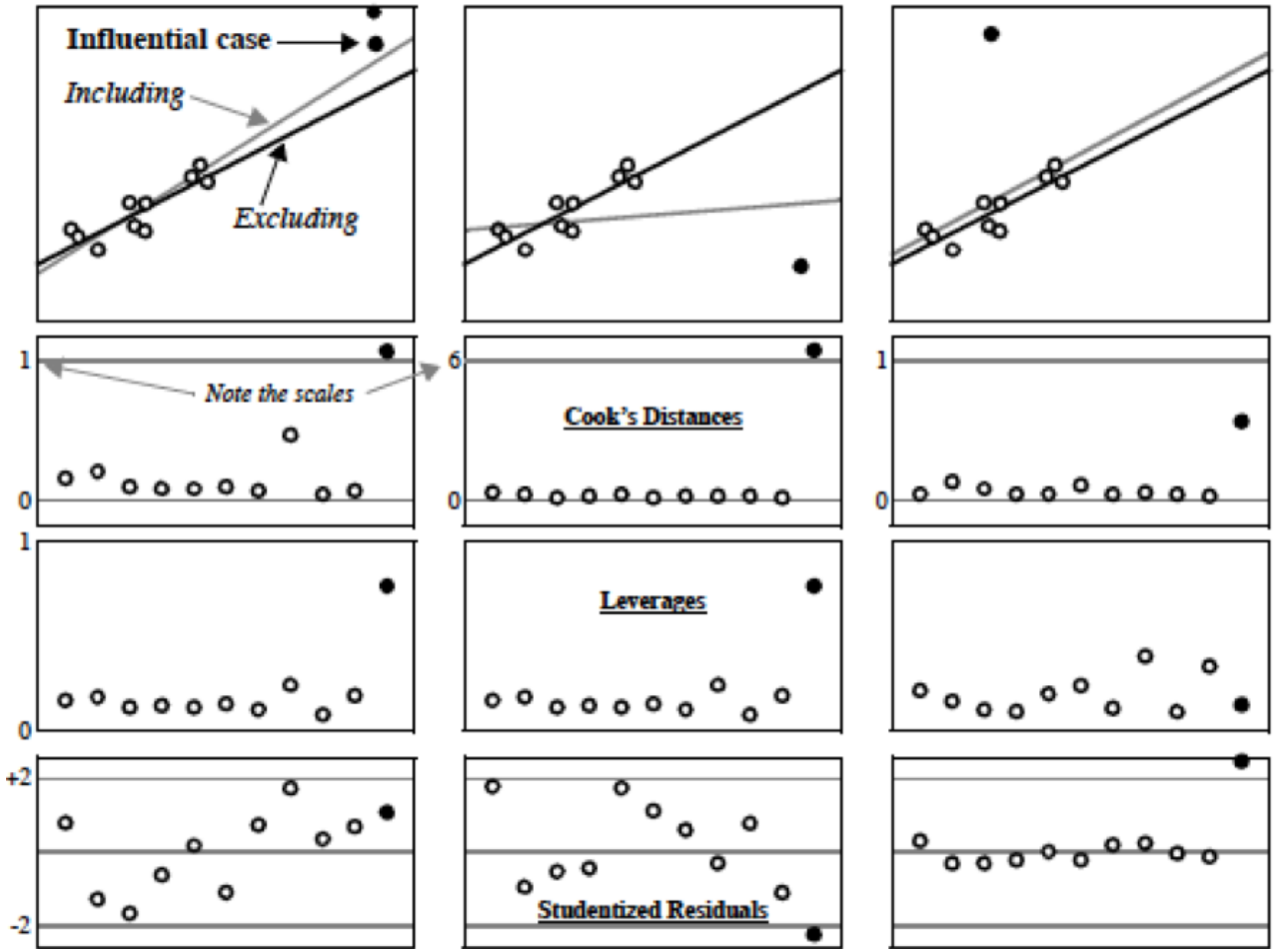
Cook's Distance: The **Cook's Distance** of a point is a measure of how much that point changes the estimated regression coefficients (and therefore the estimated fitted values).

A point has a large Cook's distance if the regression coefficients when that point is excluded from the model fit are very different from the regression coefficients when that point is included.

The Cook's distance for point i is often denoted by D_i.

There are no bounds on possible values for Cook's distance, but typical recommendations are that points with $D_i > 1$ or $D_i > 4/n$ are worth further investigation.

Comparing Influence Measures



Case Influence Statistics in R

The functions `hatvalues()`, `rstudent()` and `cooks.distance()` can be applied to a `lm` object to find the leverage, Studentized residual and Cook's Distance for each observation.

For example, with the Meadowfoam data,

```
fit_meadowfoam <- lm(Flowers ~ Intensity + Timing +
   Intensity:Timing, data = case0901)

hatvalues(fit_meadowfoam)
rstudent(fit_meadowfoam)
cooks.distance(fit_meadowfoam)
```

Most commonly, the case influence statistics are plotted against observation number, to help visually identify unusual points.

Recommendations for influential points

Display 11.8 **p. 314**

A strategy for dealing with suspected influential cases