Lecture 13: Identifying unusual observations

In lecture 12, we learned how to investigate *variables*. Now we learn how to investigate *cases*.

Goal: Find unusual cases that might be mistakes, or that might strongly influence results.

3 types of unusual cases:

1. Cases with *high leverage* have one or more extreme explanatory variable values. (Unusual X values)

2. *Outliers* do not fit the trend of the rest of the data, identified by having large residuals. (Unusual Y values)

3. *Influential cases* have a strong impact on some aspect of the regression – predicted values, R^2, test results, etc. Outliers and high leverage cases *might* be influential.

How to Identify Unusual Cases

Easy to do visually in simple linear regression, but need numerical measures to find them in multiple regression.

Identifying high leverage cases:

Definition: For simple linear regression, the *leverage* or *hat value* for case *i* is

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SSX}$$

Notes (for simple linear regression only)

1. $\sum_{i=1}^{n} h_i = 2$. (Show why on board.)

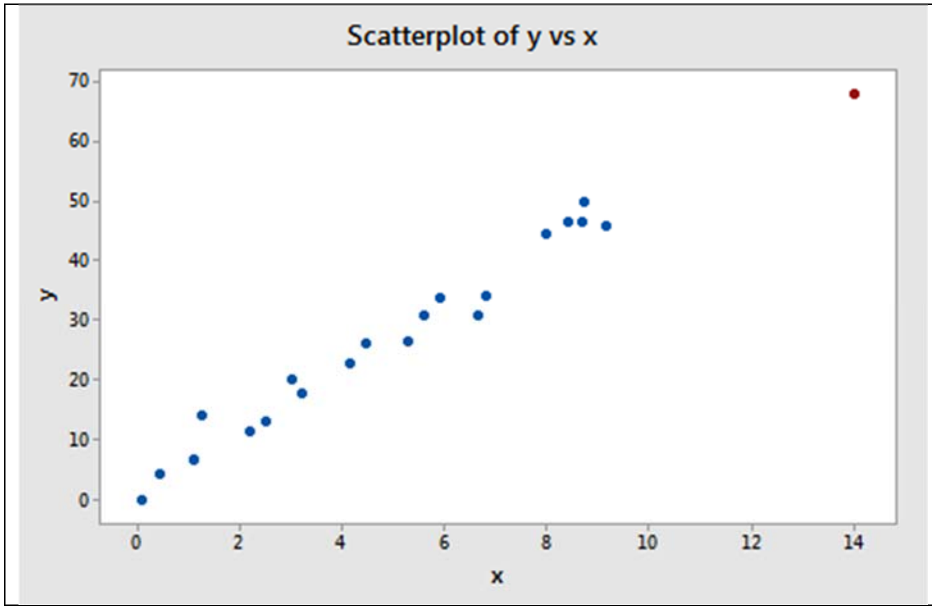
2. From #1, clearly the average is $\bar{h} = \frac{2}{n}$. We will identify:

- *high leverage* cases as those with $h_i > 2\bar{h}$, so $h_i > 4/n$
- *extremely high leverage* cases as those with $h_i > 6/n$

3. Leverage depends on the *x* values only, not the *y* values.

EXAMPLE 1: A high leverage case (simple linear regression)

(This and the next few examples are from Penn State online regression course)



$n = 21$

High leverage is 4/21 = 0.19

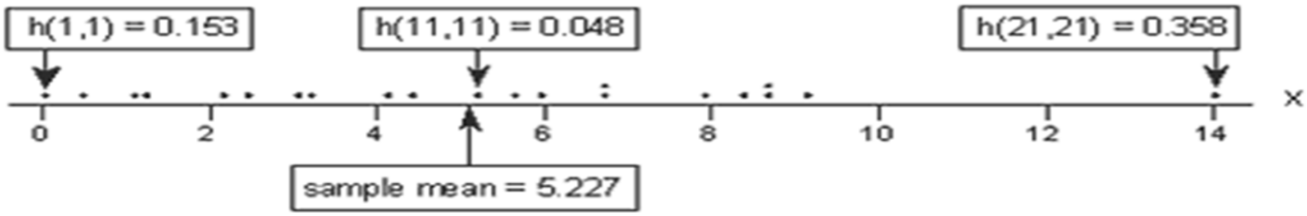
Extreme leverage is 6/21 = 0.29

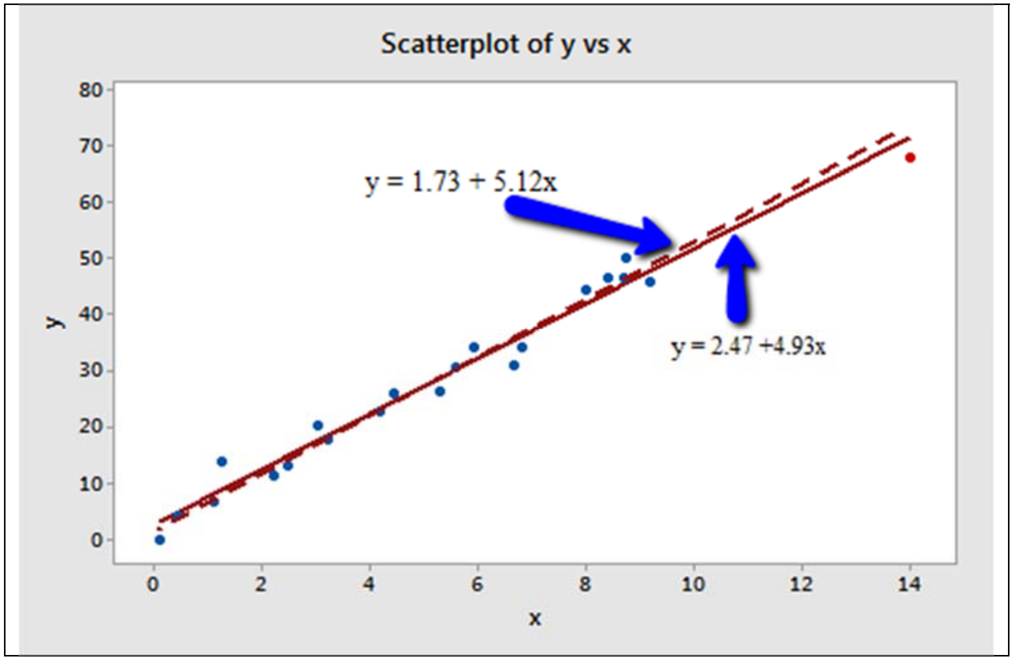
Leverage for red point is 0.36

Extreme leverage!

But is it influential?

Leverage for the x values, with them displayed on the x axis only:





Measure	With high leverage case	Without it
R^2-adj.	97.62	97.17
$\hat{\sigma}_\varepsilon = \sqrt{MSE}$	2.7	2.6
Estimated slope $\hat{\beta}_1$	4.93	5.12
s.e.($\hat{\beta}_1$)	0.172	0.200

So the case is <u>not</u> influential, even though it has high leverage.

Leverage for Multiple Regression

- Now it's the *combination* of x values for Case i that determine its leverage.
- No longer easy to write the formula (unless we use matrices)
- Idea remains the same; high values of h_i indicate large distance from other points for the combination of x values for that case.
- With k explanatory variables (so $k + 1$ coefficients), the sum of the h_i values is $(k + 1)$, so the average is $(k + 1)/n$.
- *high leverage* cases are those with $h_i > 2\bar{h}$, so $h_i > 2(k + 1)/n$
- *extremely high leverage* cases are those with $h_i > 3(k + 1)/n$
- Leverage still depends *only* on the x values, not the y values.

More Notes about Leverage for Simple *and* Multiple Regression

- $0 \leq h_i \leq 1$, always
- $Variance(e_i) = \sigma^2(1 - h_i)$ for the residuals $e_i = Y_i - \hat{Y}_i$
- $Variance(\hat{Y}_i) = \sigma^2(h_i)$
- So, *large* h_i means that case has a *small* variance on the residual and a *large* variance on the predicted value. $\hat{Y}_i$.
- Interpretation of the above: for the same set of *x* values, in repeated sampling of new *y* values, at an *x* combination with high leverage $\hat{Y}_i$ will change a lot, but the residuals will be small.
- Can picture this for linear regression – the line will come close to the *y* value at that *x*, so the residual will be small.
- <u>Estimate</u> of $Var(e_i) = MSE(1 - h_i)$
- <u>Estimate</u> of $Var(\hat{Y}_i) = MSE(h_i)$

OUTLIERS (Unusual Y values)

Identify using standardized and studentized residuals.

For Case *i*:

<u>Standardized residual</u> for Case *i* = stdres$_i$

$$= \frac{(e_i - 0)}{s.e.(e_i)} = \frac{(Y_i - \hat{Y}_i)}{\sqrt{MSE(1 - h_i)}}$$

<u>Studentized residual</u> for Case *i* = stures$_i$

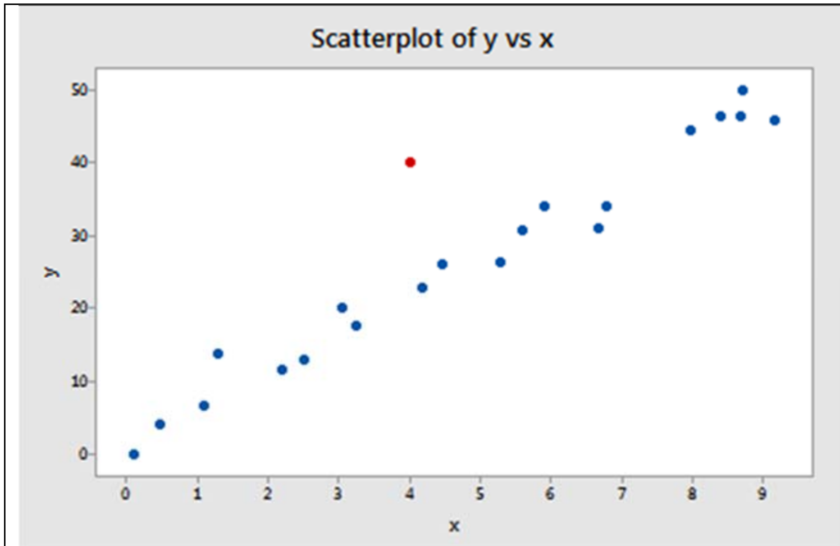
$$= \frac{(e_i - 0)}{s.e.(e_i)} = \frac{(Y_i - \hat{Y}_i)}{\sqrt{MSE_{(i)}(1 - h_i)}}$$

where $MSE_{(i)} = MSE$ for the model fit <u>without</u> Case *i*.

NOTE: Some sources define this using $\hat{Y}_{i(i)}$ as the predicted value, i.e fit for the model without Case *i*. Others call that the Studentized deleted residual.

- Moderate outliers: Cases with absolute value of either of these > 2
- Extreme outliers: Cases with absolute value of either of these > 3

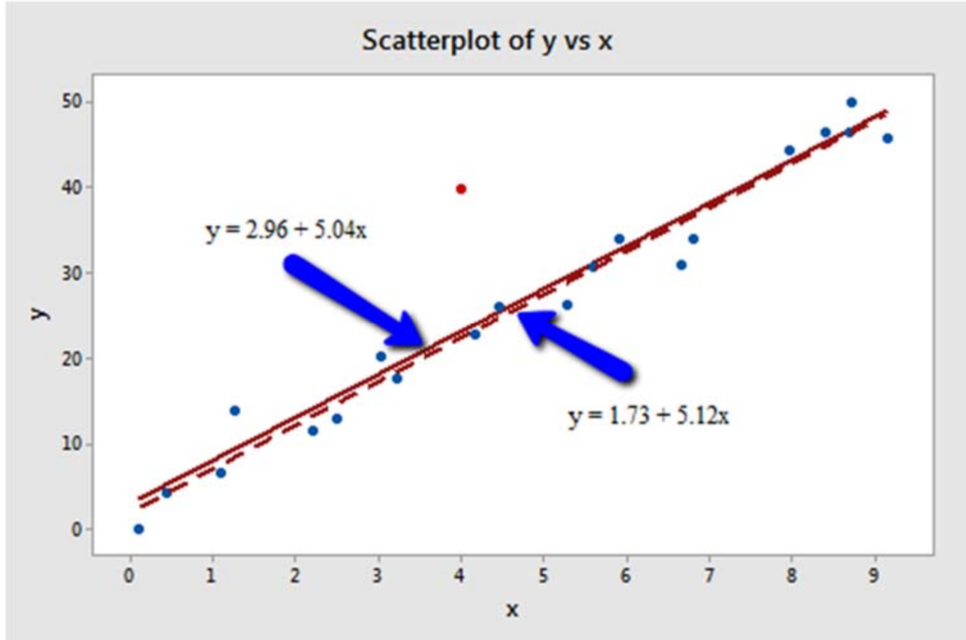
EXAMPLE 2: Outlier



rstandard = 3.68
rstudent = 6.69

So the red point is clearly identified as an extreme outlier.

Is it influential?



Measure	With outlier case	Without it
R^2-adj.	90.13	97.17
$\hat{\sigma}_\varepsilon = \sqrt{MSE}$	4.7	2.6
Estimated slope $\hat{\beta}_1$	5.04	5.12
s.e.($\hat{\beta}_1$)	0.363	0.200

It barely changes the regression equation, but variability is reduced when it is removed, as would be expected!

New Measure, Combining Both Ideas

Cook's distance combines leverage and outlier measures.

$$D_i = \frac{1}{(k+1)}(stdres_i)^2(\frac{h_i}{1-h_i})$$

Large Cook's distance implies large *stdres* or large leverage or both.

"Flag" (i.e. identify) cases with Cook's distance > 0.5 for moderate, or > 1 for extreme.

EXAMPLE 1: Cook's distance for the high leverage point is 0.702.

EXAMPLE 2: Cook's distance for the outlier is 0.36.

Another version of the formula (not in book), easier to see why it works:

Define $\hat{Y}_{j(i)}$ = predicted Y_j using model without Case i.

In other words:

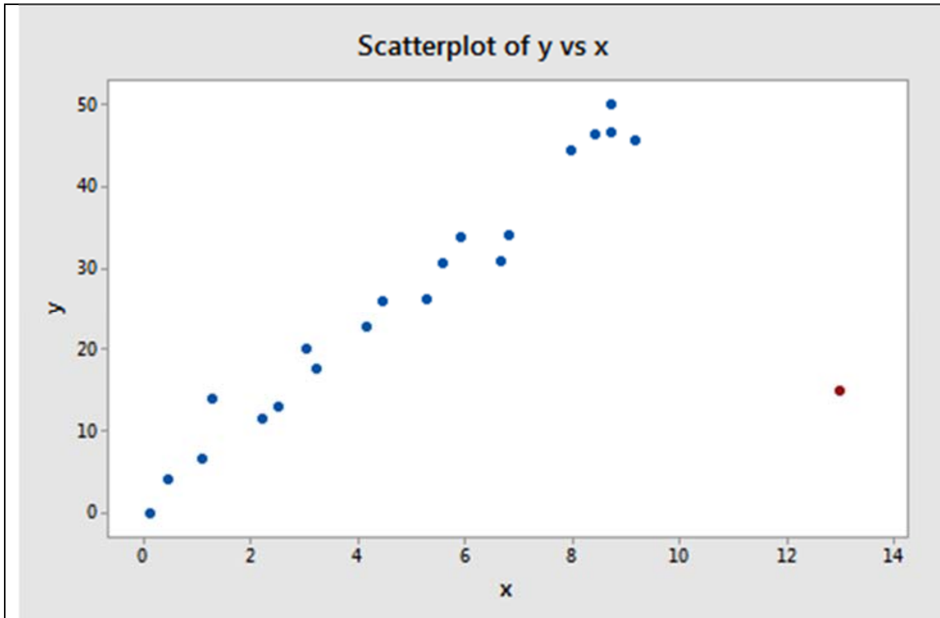
- Remove case i
- Fit model
- Use it to predict all of the *other* cases, $j = 1, \ldots, n$

Then

$$D_i = \left(\frac{1}{k+1}\right) \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{MSE}$$

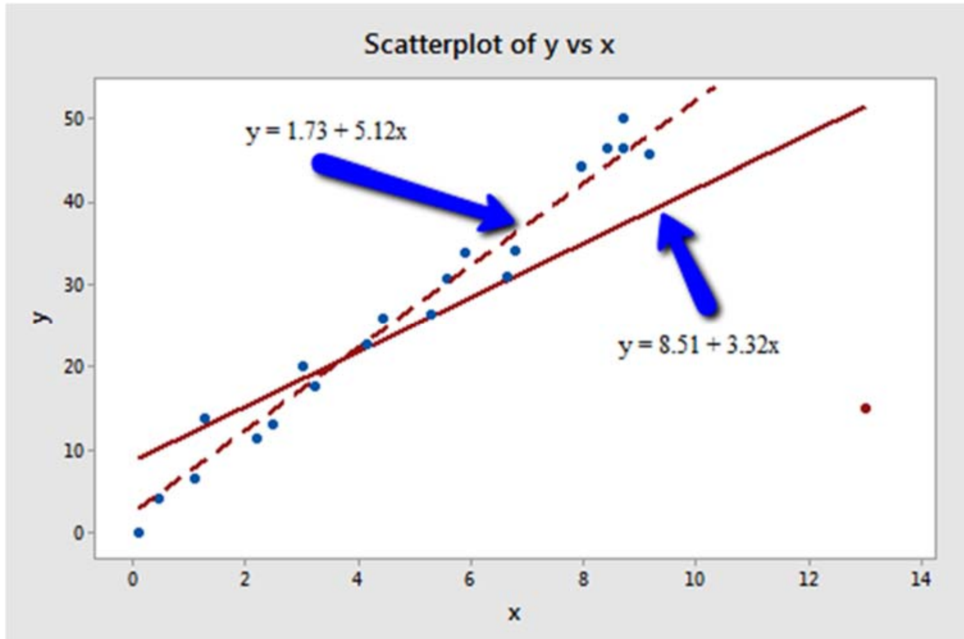
It's the "distance" (squared and normalized) between the predicted values for all cases, using the model *with* Case i included, and the model *without* Case i included.

EXAMPLE 3:



This point has:
Leverage = 0.31
Std. residual = -4.23
Cook’s D = 4.05
All extreme!

Let’s see what happens when it’s removed.



Measure	With case	Without it
R^2-adj.	52.84	97.17
$\hat{\sigma}_\varepsilon = \sqrt{MSE}$	10.4	2.6
Estimated slope $\hat{\beta}_1$	3.32	5.12
s.e.($\hat{\beta}_1$)	0.686	0.200

NEXT: Diagnostics in R, then Real estate example.