OUTLIERS

Abdullah Farouk

March 2018

What is an Outlier?

It is an observation that is VERY different from the others in your data

Some potential causes are:
– Error in recording the measurement
– Failure of the measurement process/tool
– Representative of the population sampled

Is it a Spurious data point?

They are data points whose value do not teach us much about the subject matter of interest

We remove spurious data without guilt

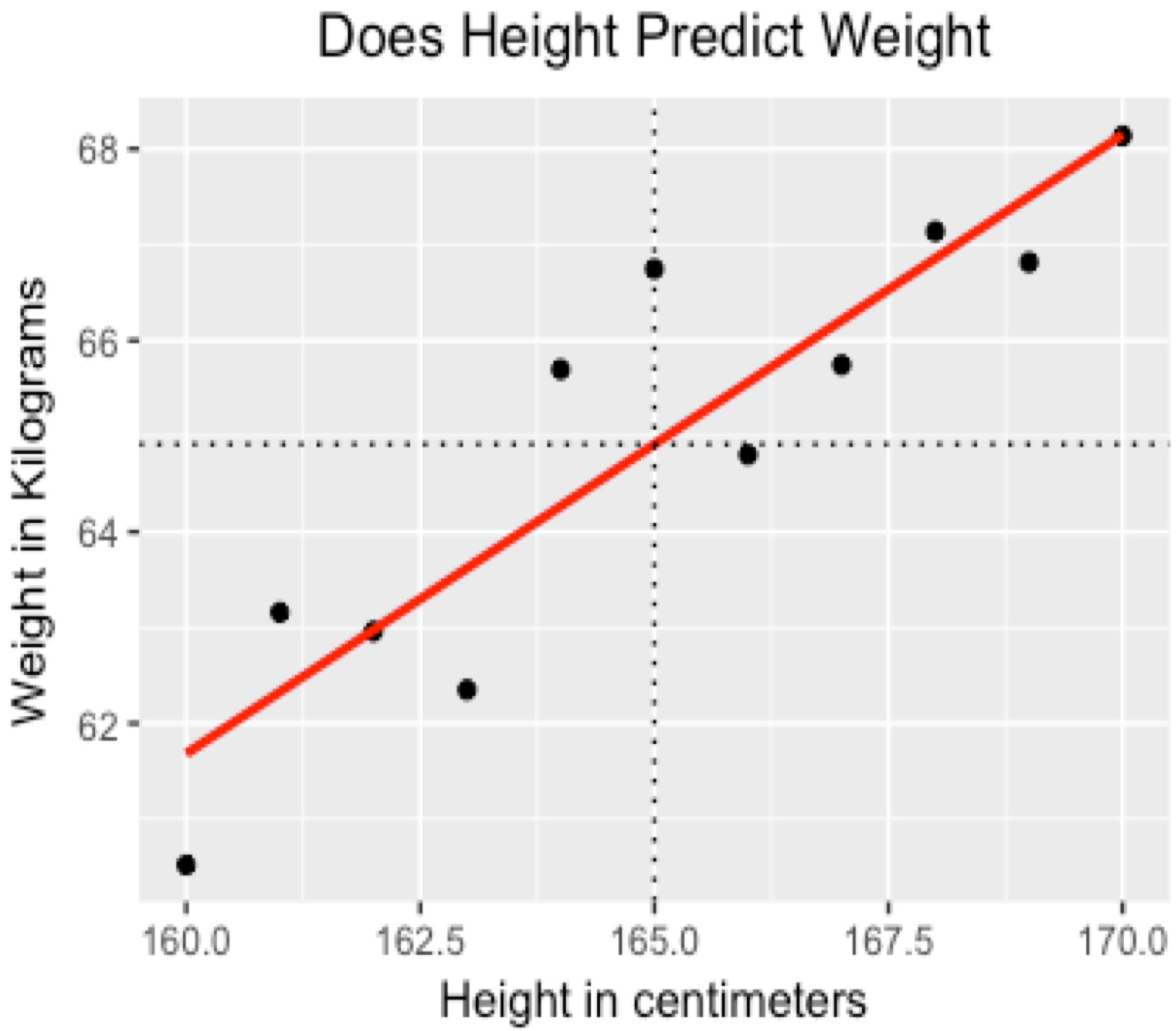
Not all outliers are spurious!

A Simple Example

Can the height of an individual predict their weight?

W = -45 + 2H/3 + ε, $\varepsilon \sim$ N(0,1) iid

Observations on 11 individuals.



Lets introduce some outliers

In a regression model, they are data points with an extreme response variable (Y)

W_6 = W_6 - 3Var(e)

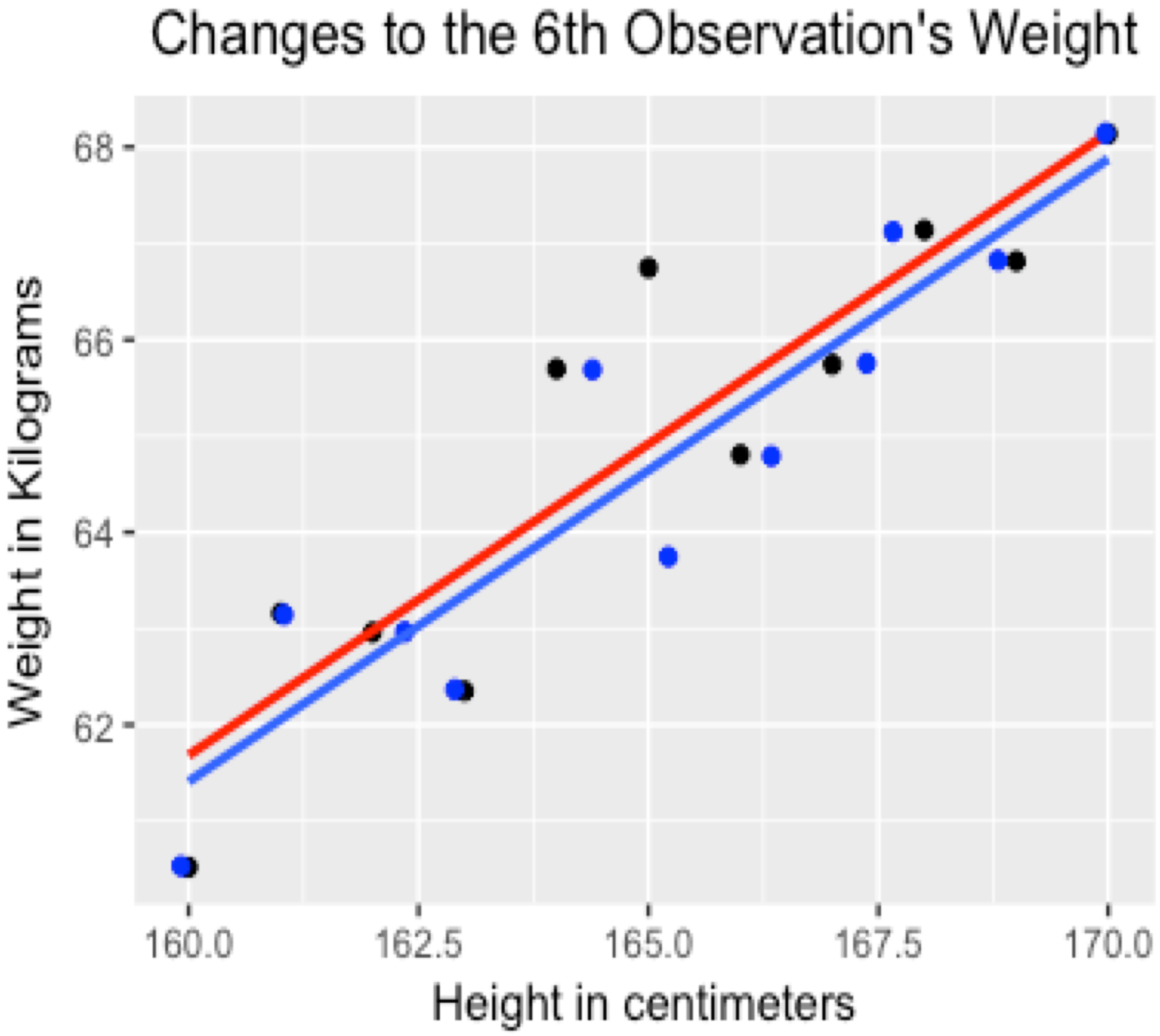
W_{11} = W_{11} - 3Var(e)

What do you think will happen to the line of best fit?

Changing the Weight of the 6th observation

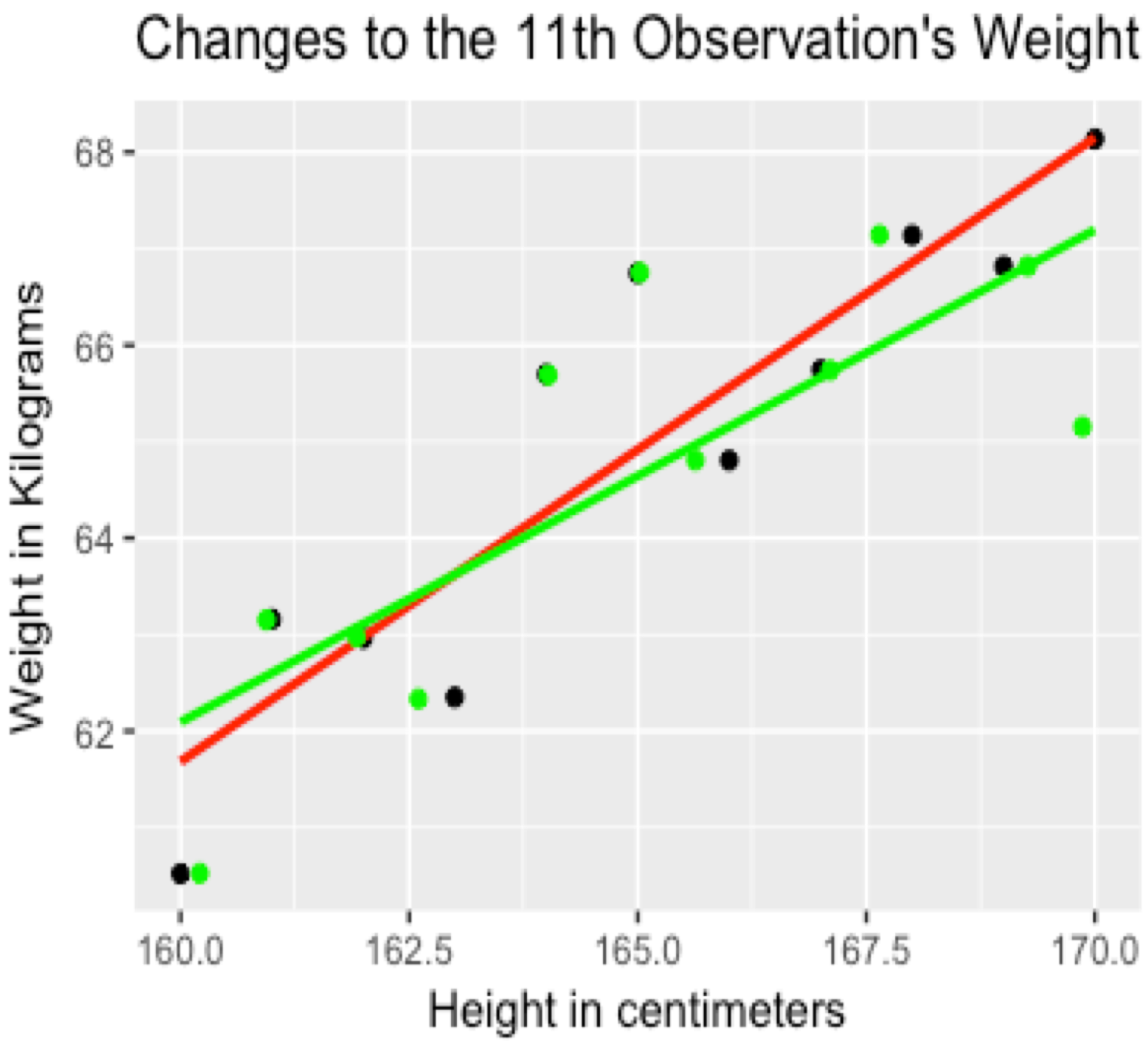
- Red line – fit without outlier

- Blue line – fit with outlier

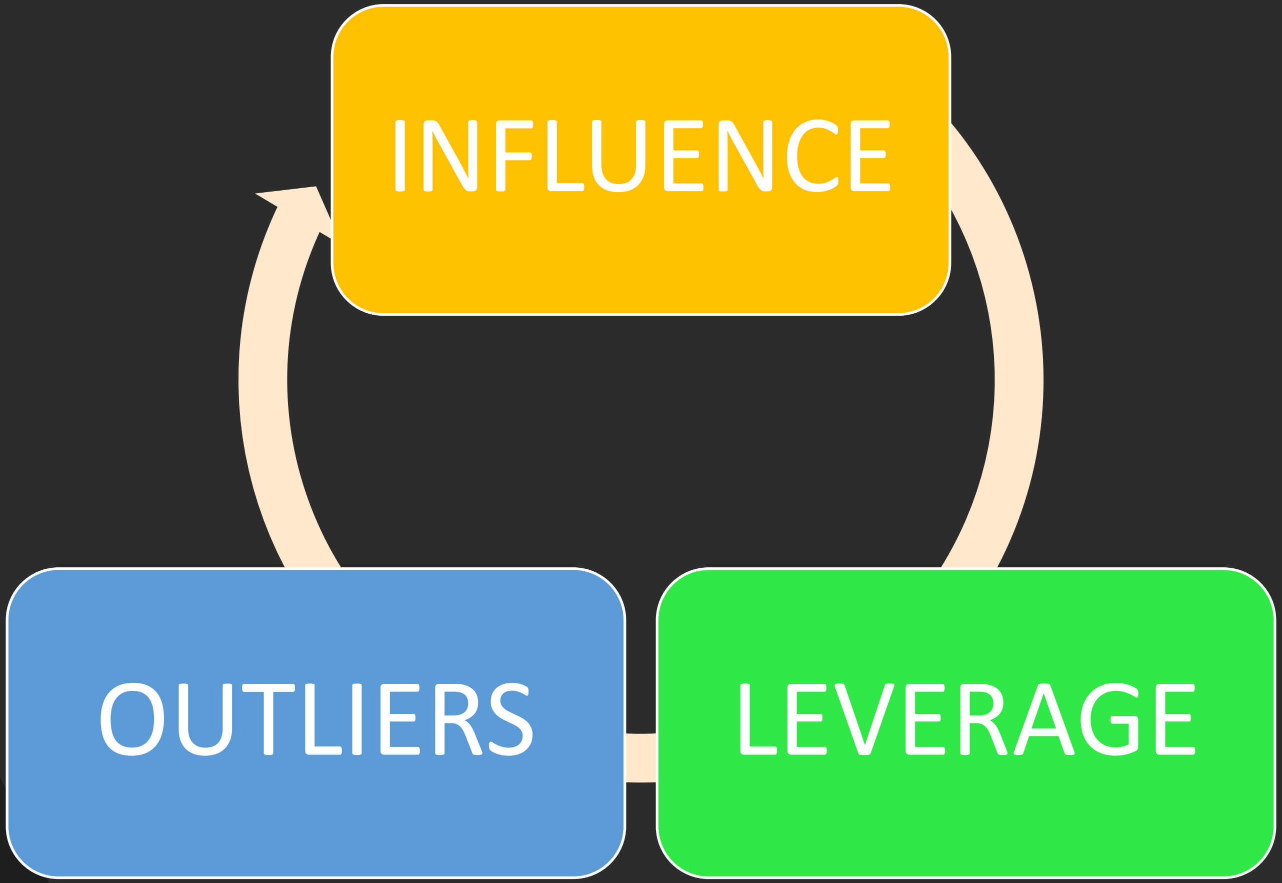


Changing the Weight of the 11th observation

- Red line – fit without outlier
- Green line – fit with outlier



Why Did this Happen?



INFLUENCE

OUTLIERS

LEVERAGE

Leverage

- Distance between x_i and its mean ($\overline{x}$)
- Observations further away from the mean have higher leverage

Influence

Data points with an extreme Y and an extreme X value are influential.

Why are these points influential?

Removing them leads to significant changes in regression results

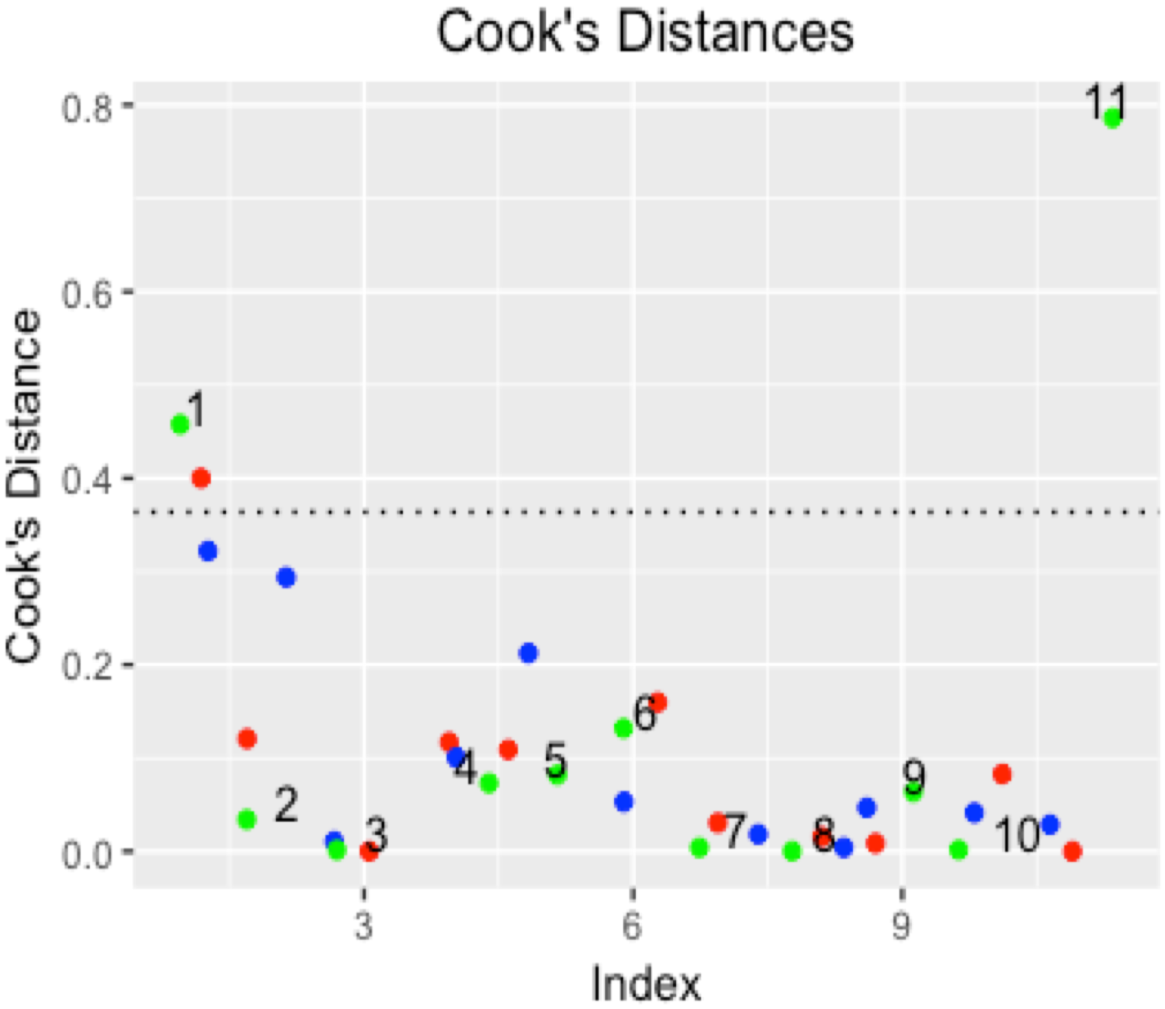
Diagnostic Measures

Cook’s Distance

Residual Plots

Cook’s Distance (CD)

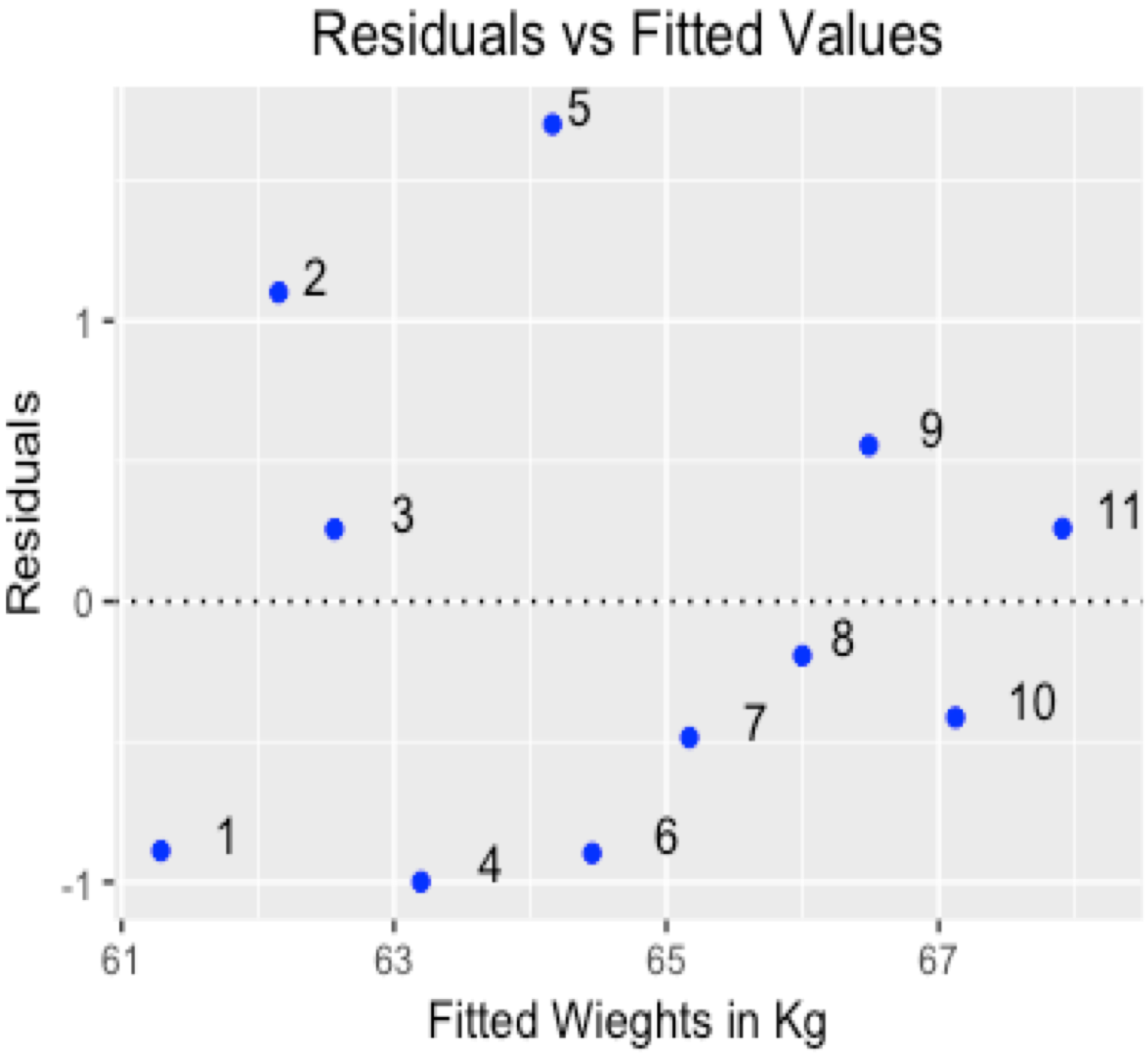
- Deletes the i^{th} data point and looks at the difference in predicted y –values
- Observations with a CD value > 1 (small samples) and > 4/n (large sample) are influential
- The 11th observation is very INFLUENTIAL



Red dots – No outliers
Blue dots – Observation 6 is an outlier
Green dots – Observation 11 is an outlier

Residual Plots

- What is this plot supposed to show?
- Blue dots - observation 6 is an outlier
- Outliers (extreme y values) with low leverage:
    - Appear as large residuals near the center of the plot



How can you deal with Outliers?

1

Try and identify it’s causes

2

Conduct your analysis with and without it

3

Use Robust statistics

References

- Altman N, Krzywinski M. Points of Significance: Analyzing outliers: influential or nuisance?.
- http://www.lithoguru.com/scientist/statistics/course.html

Thank you for Listening!

Why are outliers important?

Goal is to make accurate predictions

Outlying data can help test the stability of our predictions

If removing them severely alters the fit or sways the outcome of inference, a more complete model may be needed.

How to calculate leverage?

Recall the Hat matrix

$$\hat{Y} = \mathrm{HY}, \qquad \mathrm{H} = \mathrm{X}(X^T X)^{-1} X^T$$

- The leverage of each predictor variable is given by the diagonals of this matrix h_{ii}

$$0 \leq h_{ii} \leq 1$$

- If $h_{ii} > \frac{2P+2}{n}$ → high leverage

Cook’s Distance

- Deletes the i^{th} data point and looks at the difference in predicted y –values. It is calculated as:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_{j_{(i)}} - \hat{Y}_j)^2}{P{s_e}^2}$$

$$D_i = \frac{{sr_i}^2 \times h_{ii}}{P(1 - h_{ii})}$$

Note: P = # of predictors and sr_i = Studentized Residuals