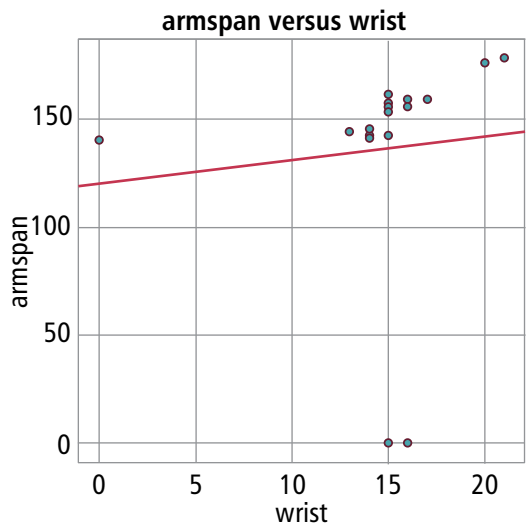
# Removing outliers using iNZight

## Relating to page 30 of *Level 3 Bivariate Data Learning Workbook*

The following data gives the armspan and wrist measurements (in centimetres) for a random sample of 17 students from the New Zealand CensusAtSchool database.

| | armspan | wrist |
|---|---|---|
| 1 | 156 | 16 |
| 2 | 161 | 15 |
| 3 | 178 | 21 |
| 4 | 144 | 13 |
| 5 | 159 | 16 |
| 6 | 159 | 17 |
| 7 | 142 | 14 |
| 8 | 144 | 14 |
| 9 | 0 | 15 |
| 10 | 155 | 16 |
| 11 | 156 | 15 |
| 12 | 140 | 0 |
| 13 | 176 | 20 |
| 14 | 142 | 15 |
| 15 | 145 | 14 |
| 16 | 154 | 15 |
| 17 | 0 | 16 |

iNZight is used to produce a scatter graph, with a linear trend line to show the relationship between wrist circumference and armspan, for this sample. Use **Get Summary** to produce the equation of the trend line, and the correlation coefficient.



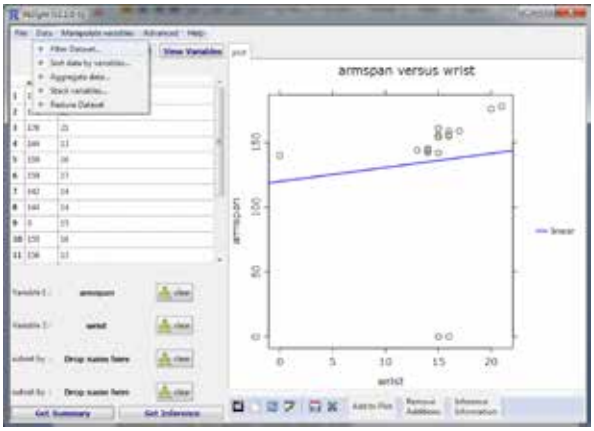

```
Summary for gender = female:
---------------------------------
Linear trend:

      armspan = 119.9 + 1.081 * wrist
      Linear correlation = 0.09
```

Visually, the linear trend line that has been fitted to the data does not appear to be a good model for the trend. This is confirmed by the correlation coefficient of 0.09 for the trend line.

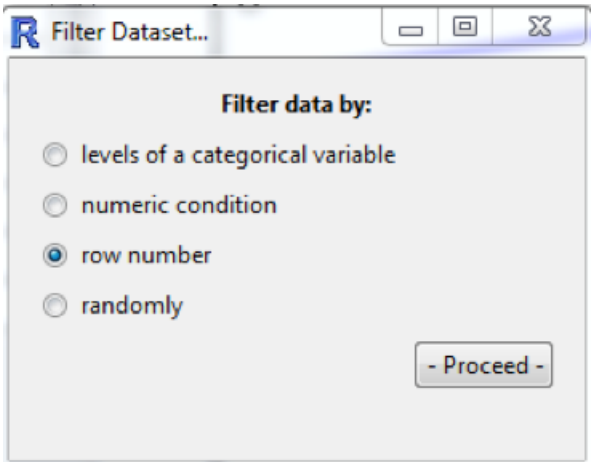
On examining the data it can be seen that two points represent students with armspans of length 0 cm (rows 9 and 17), and one other point represents a student with a wrist circumference of 0 cm (row 12). It is evident that the data values with wrist values and armspans of zero should be removed.

To do this on iNZight, follow the steps below:

- Click on **Data**
- Select **Filter Dataset**

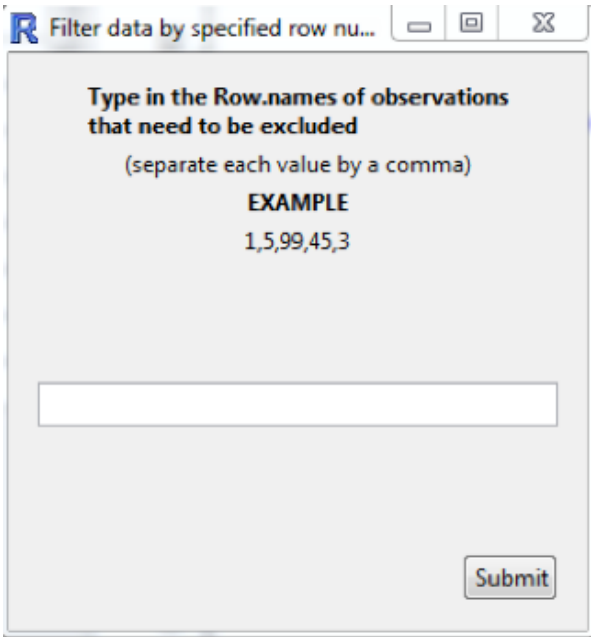



This gives the following screen:

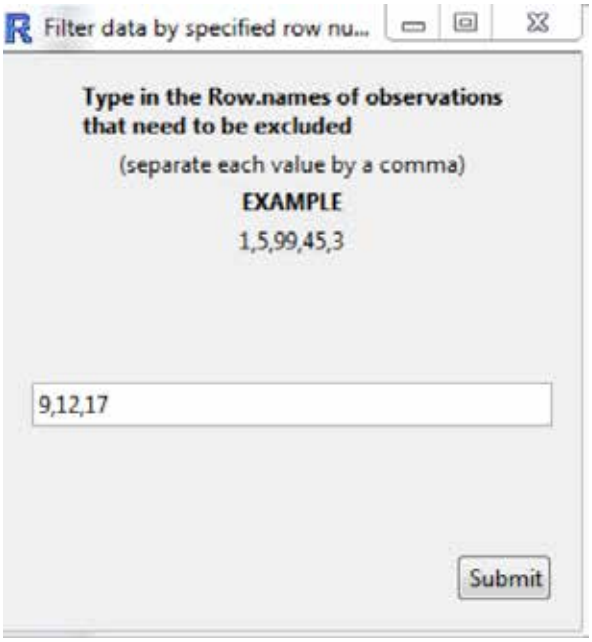



- Select **row number**
- Click on **Proceed**

The following screen appears.



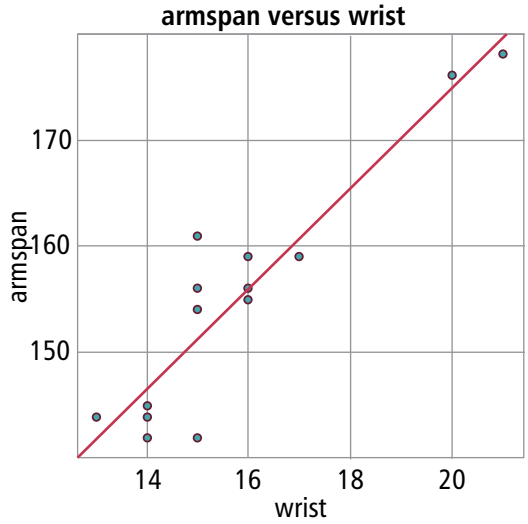
- Type in the row numbers, separated by commas, for data that is to be removed.




- Click on **Submit**

The data set now has these three rows removed (these row numbers are now missing from the data set).

A new scatterplot with trend line can then be produced as before. This is shown below.




Use **Get Summary** to produce the equation of the new trend line, and the correlation coefficient.

Your screen should now show the following.

```
Summary of armspan versus wrist:
---------------------------------------
Linear trend:

       armspan = 81.04 + 4.69 * wrist
       Linear correlation = 0.92
```

Again, notice how much better the line fits the points once the outliers are removed, with reduced scatter about the trend line and a correlation coefficient that is much closer to 1.

**Note:** To return the data set to its original form (before the outliers were removed), press **Data**, then select **Restore Dataset**