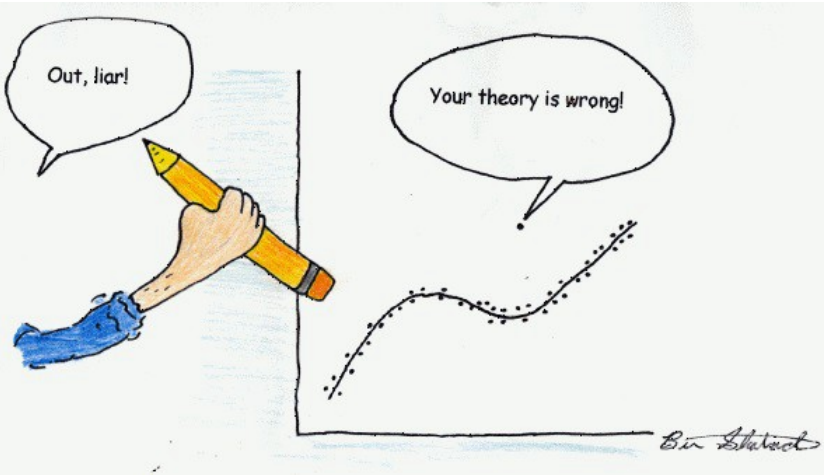
Robust Statistics

Advanced Statistics

2019-03-05

Outliers

Really singular observations that can ruin our model



Robust Statistics: Strategies to deal with outliers

- Use methods that are not sensible to them
  - Leaving some observations out of the estimation
  - Weighting the observations

Reminder on how to read regression results

```
x = runif(100)
y = 2 * x + 3 + rnorm(100, 0, 0.1)

summary(lm(y~x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25850 -0.05758 -0.01924  0.07377  0.26323
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.01408    0.02100  143.51   <2e-16 ***
## x            1.99770    0.03609   55.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1045 on 98 degrees of freedom
## Multiple R-squared:  0.969,    Adjusted R-squared:  0.9687
## F-statistic:  3063 on 1 and 98 DF,  p-value: < 2.2e-16
```

Exercise 1: Summary statistics

Check that for the following data generated from a student's t with df = 1 the mean is not a good summary of the location of the data:

```
set.seed(123)
rt(15, 1)
```

```
##  [1]  -0.2624269  -3.0702730  -0.2721196   0.7431824  43.3592961
##  [6]  -1.9774099  -0.8539835  -0.5448942   1.0924380  -2.8547817
## [11]  -0.1757272 315.2622682   1.5399306  -5.7825676  -0.7510694
```

Propose 3 other summary statistics that are robust and verify the robustness with this data.

Exercise 2: Detect outliers

Create a toy example where univariate boxplots are not enough to detect the outliers of the data.

Exercise 3: Robust linear regression

Use the `starwars` dataset that is included in the `dplyr` package:

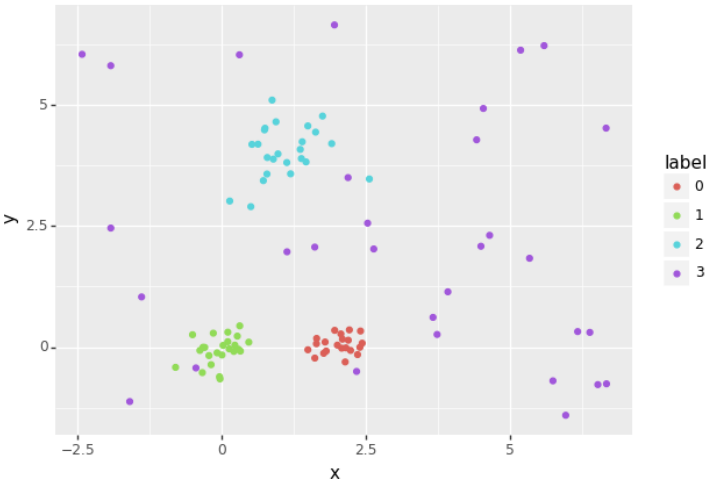
- **Briefly** describe the dataset.
- Consider the variables `height` and `mass`, plot univariate charts to study the range and other summary statistics.
- Plot mass vs height and describe it.
- Fit two different regression models. Plot and discuss the results. Useful: lmrob or MASS::rlm.
- Inspect the weights of the M-estimators. Plot with different colors the observations with small weights.
- Retrieve the name of the most extreme outlier and show a picture of him/her/it.
- Repeat the procedure excluding this outlier.
- Which would be the predicted mass for a character of height 170?

BONUS: Code your own algorithm for the robust linear model

Exercise 4: Robust EM

TClust is a robust EM algorithm similar to the trimmed mean in the sense that it leaves out of the estimation some observations that are extremes.

- Generate a mixture of normal distributions in 2 dimensions (with K>=3) and add some observations coming from a uniform in the rectangle where the distributions are made. Useful: MASS::mvrnorm.



- Apply tclust and a classical EM to cluster the generated data.
- Compare the results.