Robust Statistics

Advanced Statistics

2020-11-06

LAB 4

- Deadline: November, 18
- To: adv.statistics.2020@gmail.com.
- Subject: LAB 4
- Report file name: LAB4_LastName1_LastName2

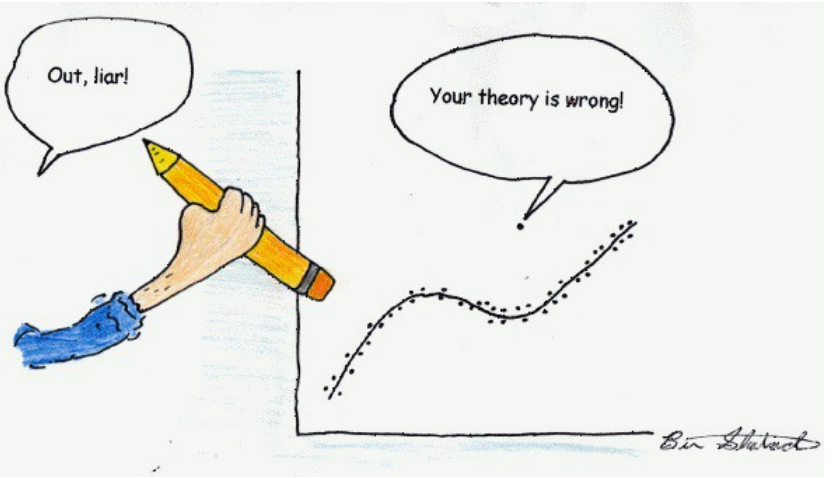
For this lab:

- You'll solve the 4 exercises on these slides
- You should submit a report based on the exercises

Guidelines: We expect a self-contained report with answers, figures and results. You can use RMarkdown, it is not mandatory. Additionally, you will send the code (.R) or the file that builds the report (.Rmd), we only want to look at it if there is a mistake on the report.

Outliers

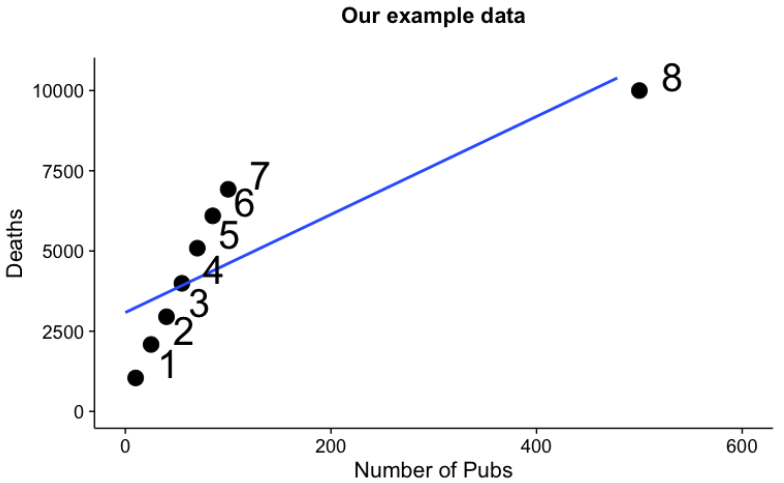
Really singular observations that can ruin our model

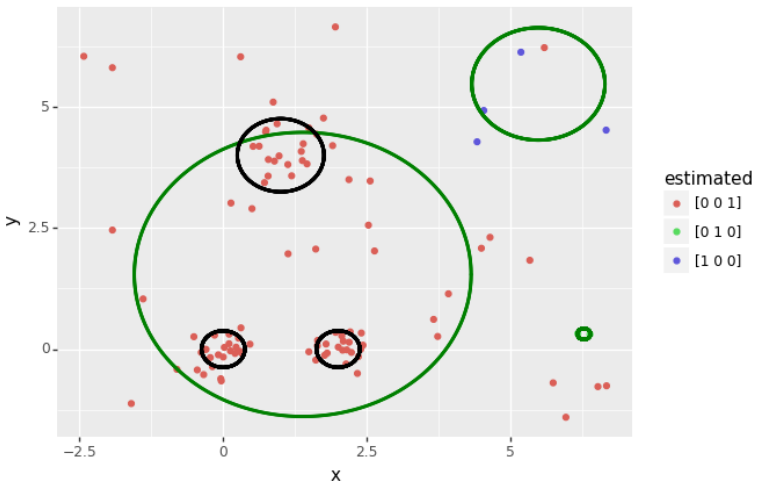


Robust Statistics: Strategies to deal with outliers

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

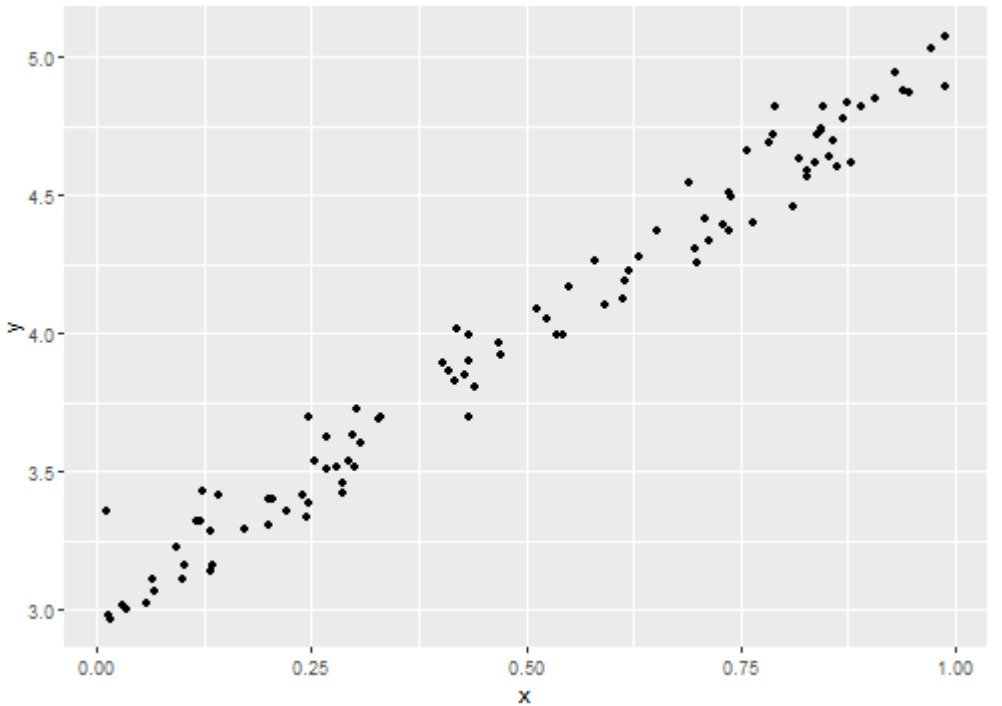
Some examples





Reminder on how to read regression results

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



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
## -0.236092 -0.072562  0.006706  0.074806  0.187934
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.02060    0.02048  147.52   <2e-16 ***
## x            1.95524    0.03720   52.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09858 on 98 degrees of freedom
## Multiple R-squared:  0.9657,    Adjusted R-squared:  0.9654
## F-statistic:  2762 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  -1.97740
##  [7]  -0.8539835  -0.5448942   1.0924380  -2.8547817  -0.1757272 315.26226
## [13]   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:

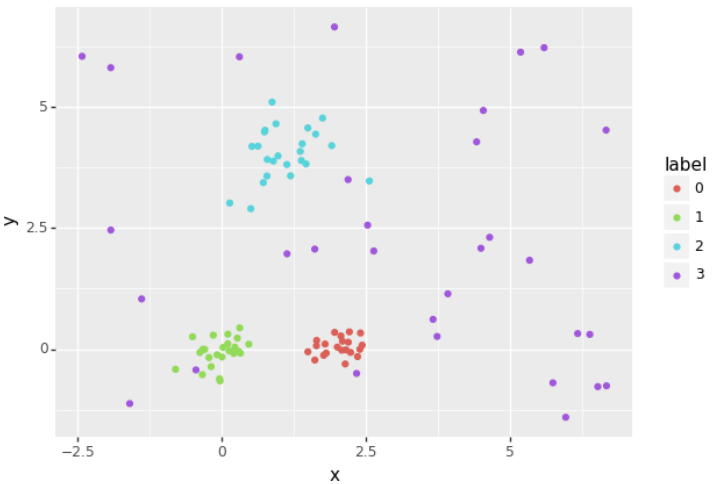
- **Briefly** describe the dataset.
- Consider the variables `height` and `mass`, plot univariate charts to study the range and other summary statistics.
- Plot mass (y-axis) vs height (x-axis) and describe it.
- Fit two different regression models. Plot and discuss the results. Useful: lmrob or MASS::rlm.
- 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/they/it.
- Repeat the procedure excluding this outlier.
- Which would be the predicted mass for a character of height 170?

BONUS (not graded): 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.