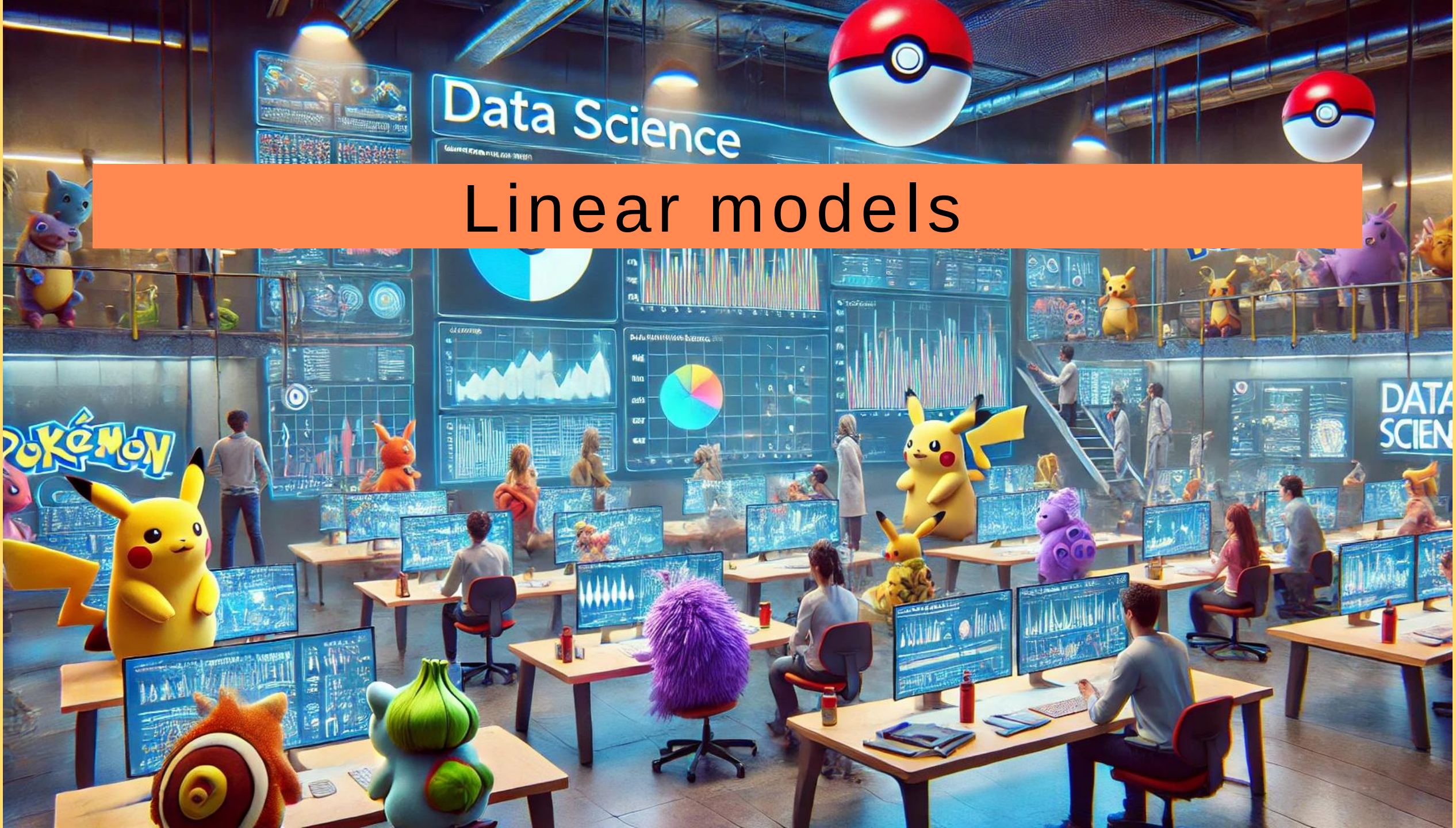
Linear models



The key stages of the linear model

Definition of the question

Exploration

Model construction

Checking application conditions

Interpretation and conclusion

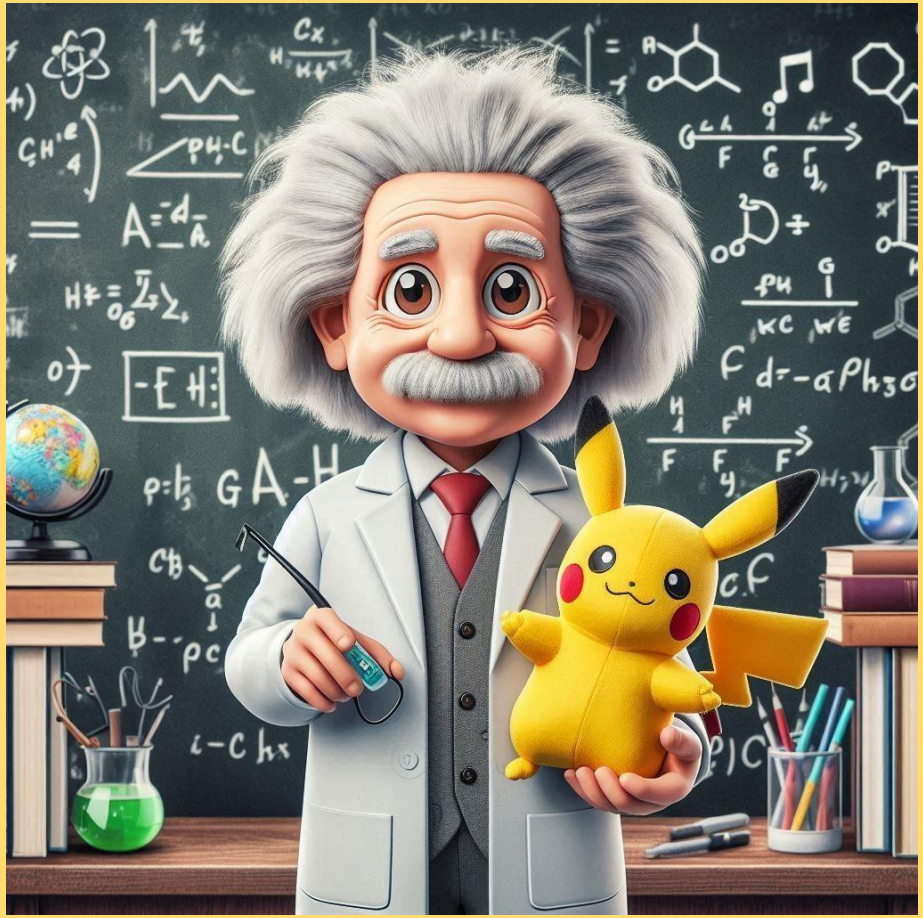
The key stages of the linear model

Definition of the question

Defining your question: the priority

- "If I had one hour to save the world, I'd spend 59 minutes defining the problem and one minute finding solutions." –

- Albert Einstein.



What make a fast Pokemon?

What make a Pokemon Happy?

Dataset

```
> summary(data2 %>% select(base_happiness,height_m,attack,type1,percentage_male))
 base_happiness      height_m          attack          type1            percentage_male
 Min.   :  0.00   Min.   : 0.200   Min.   : 10.00   Length:185         Min.   :  0.0
 1st Qu.: 70.00   1st Qu.: 0.500   1st Qu.: 55.00   Class :character   1st Qu.: 50.0
 Median : 70.00   Median : 1.000   Median : 70.00   Mode  :character   Median : 50.0
 Mean   : 63.92   Mean   : 1.221   Mean   : 74.95                      Mean   : 57.4
 3rd Qu.: 70.00   3rd Qu.: 1.500   3rd Qu.: 90.00                      3rd Qu.: 50.0
 Max.   :100.00   Max.   :14.500   Max.   :155.00                      Max.   :100.0
                  NA's   :1                                            NA's   :22
```

name	base_happiness	height_m	attack	percentage_male	speed	type1
Squirtle	70	0.5	48	88.1	43	water
Wartortle	70	1.0	63	88.1	58	water
Blastoise	70	1.6	103	88.1	78	water
Pikachu	70	0.4	55	50.0	90	electric
Raichu	70	NA	85	50.0	110	electric
Psyduck	70	0.8	52	50.0	55	water
Golduck	70	1.7	82	50.0	85	water
Poliwag	70	0.6	50	50.0	90	water
Poliwhirl	70	1.0	65	50.0	90	water
Poliwrath	70	1.3	95	50.0	70	water
Tentacool	70	0.9	40	50.0	70	water
Tentacruel	70	1.6	70	50.0	100	water

What do you want to explain?

Response variable

- Speed score

Based on what?

Where explanatory variables

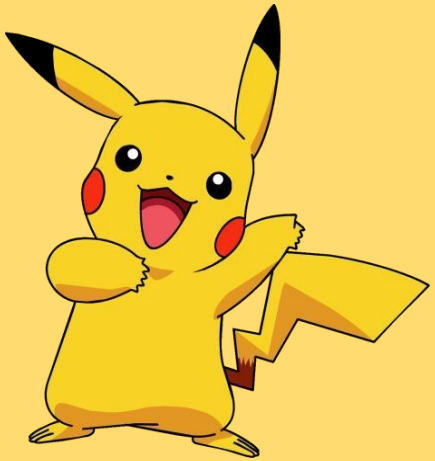
- Type of Pokemon

- Attack score

- Height (cm)

What do you want to explain?

Response variable



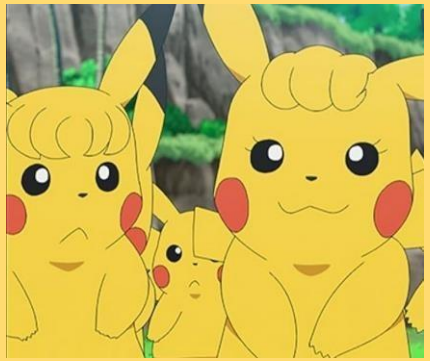
- Pokemon happiness score

Based on what?

Where explanatory variables

- Type of Pokemon

- Attack score



- Sex ratio

Different types of linear models (LM)

Simple regression / Multiple regression

quanti ~ quanti

ANOVA

quanti ~ quali

ANCOVA

quanti ~ quali + quanti

2 different variable types

- Objective: Explain XX in terms of XX and XX etc...
- The response (dependent) variable: Y
- Explanatory variables (independent): X
- Y is analyzed as a function of X: Y ~ X

The key stages of the linear model

Definition of the question

Exploration

The key stages of the linear model

Definition of the question

Exploration

Qualitative variables ← Sampling effort

Quantitative variables ← outliers

Link between the response variable and the other explanatory variables ← relationships

The key stages of the linear model

Definition of the question

Exploration

Qualitative variables ← Sampling effort

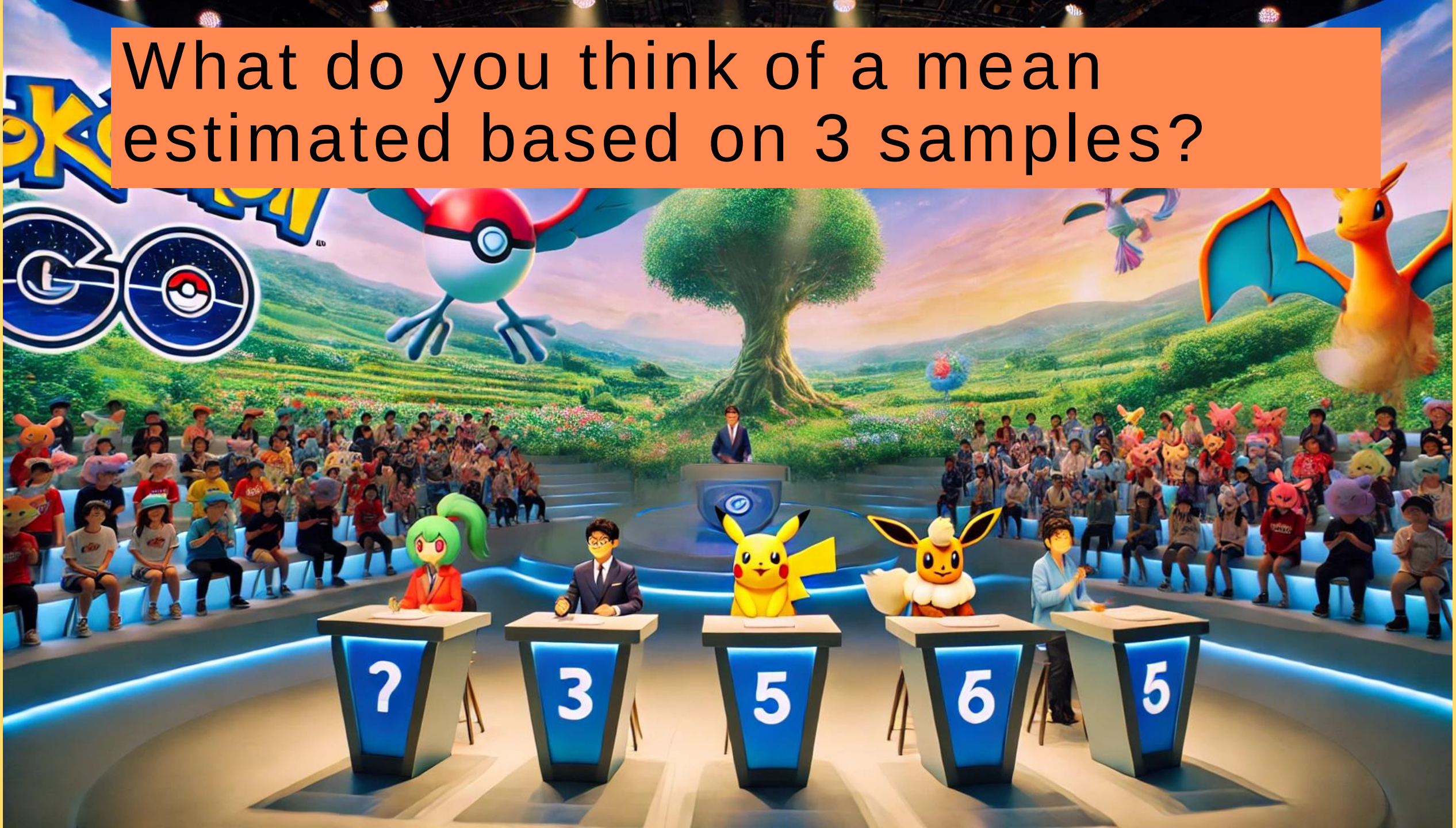
Sampling effort in type of Pokemon

Is sampling effort balanced across categories?

```
table(data2$type1)

  dark electric   flying    water
    29       39        3      114
```

=> let's remove the flying category !

What do you think of a mean estimated based on 3 samples?



The key stages of the linear model

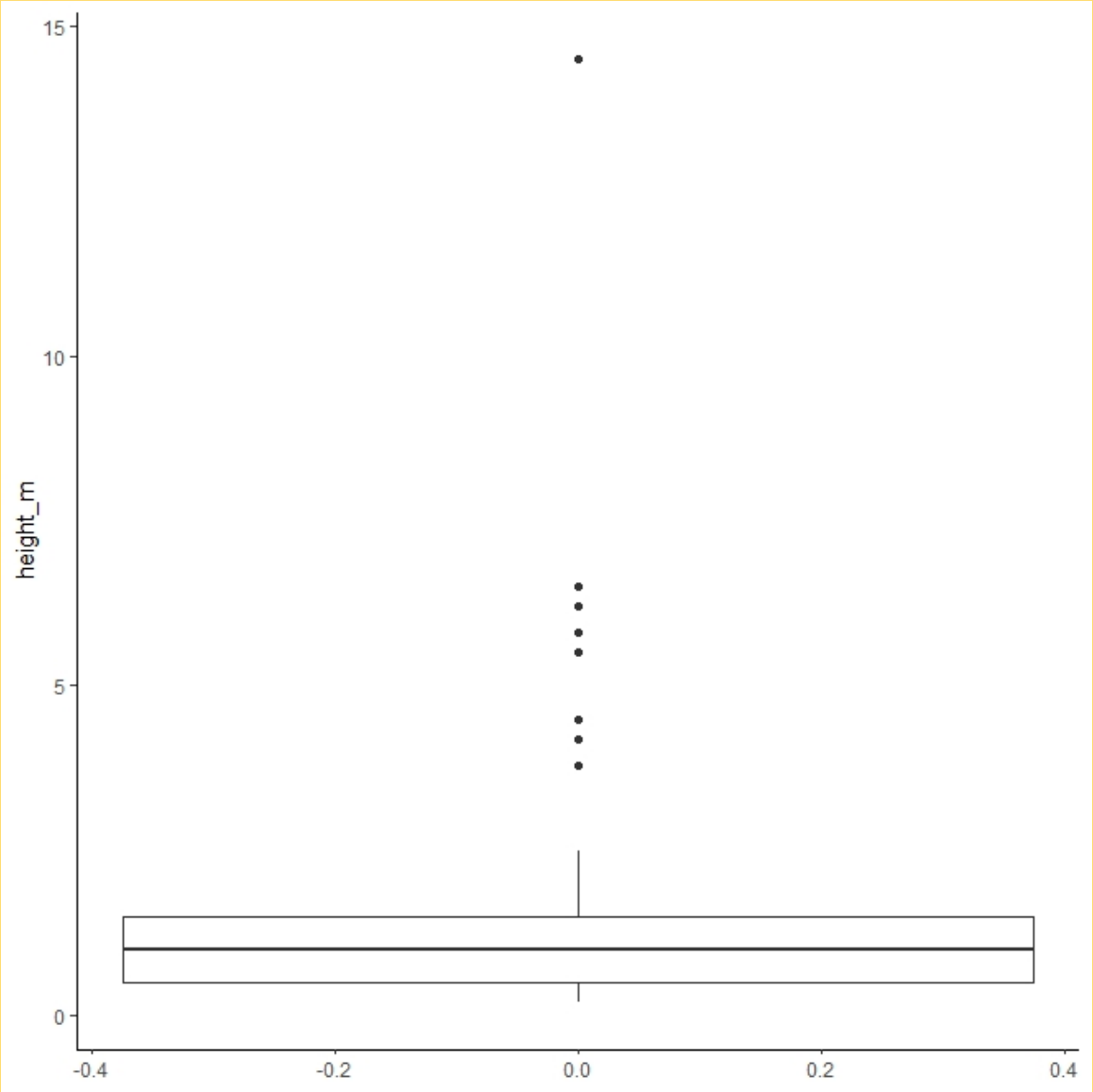
Definition of the question

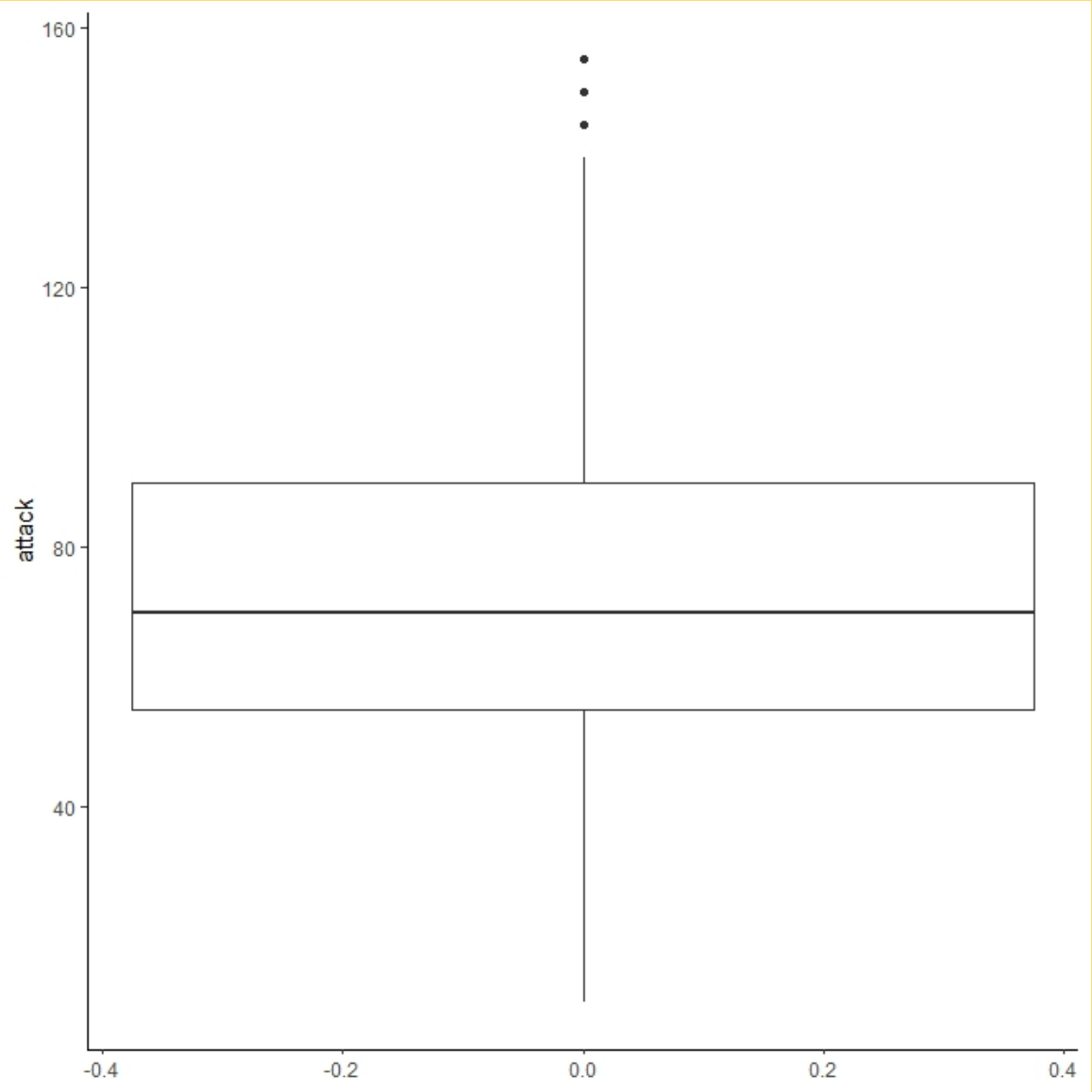
Exploration

Qualitative variables ← Sampling effort

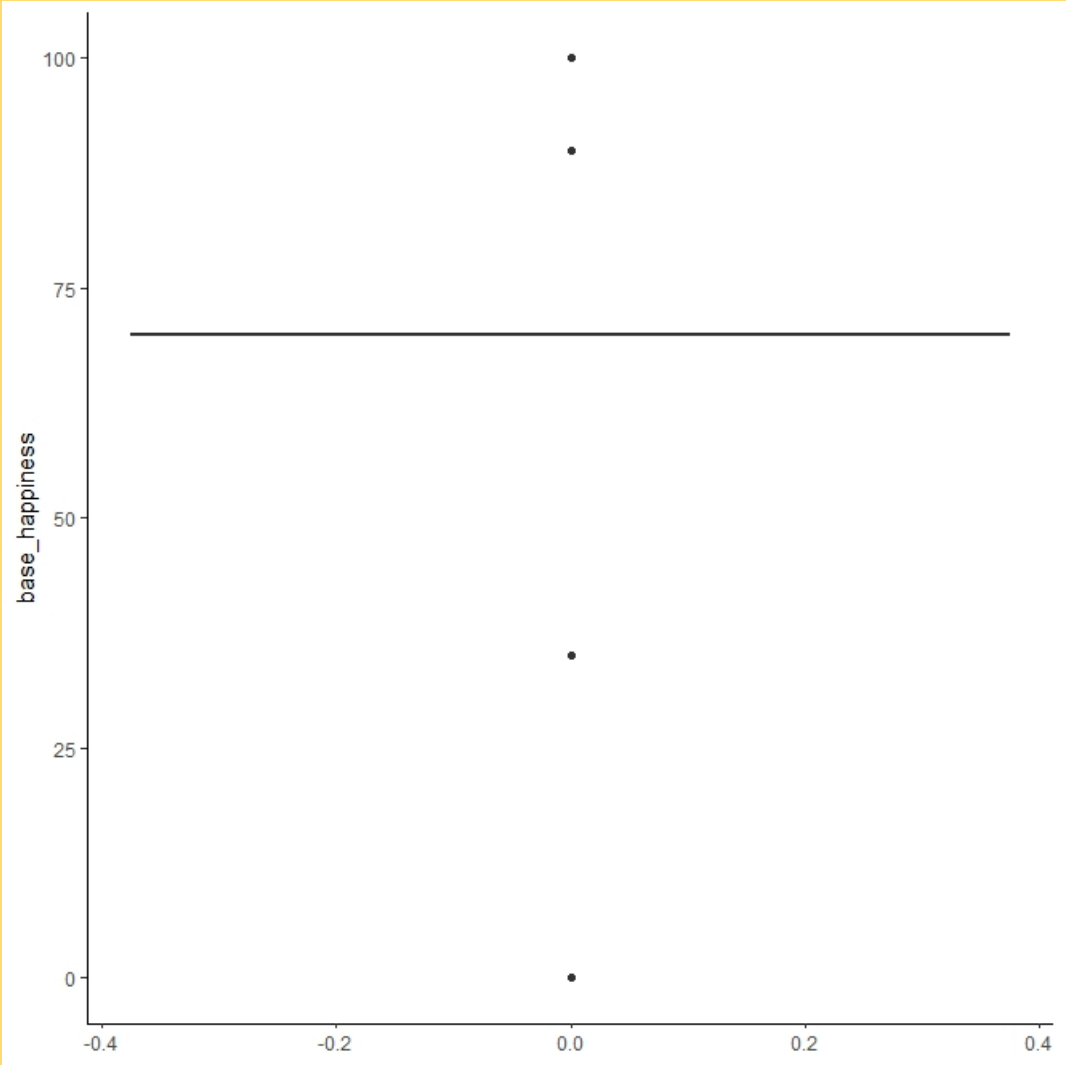
Quantitative variables ← outliers

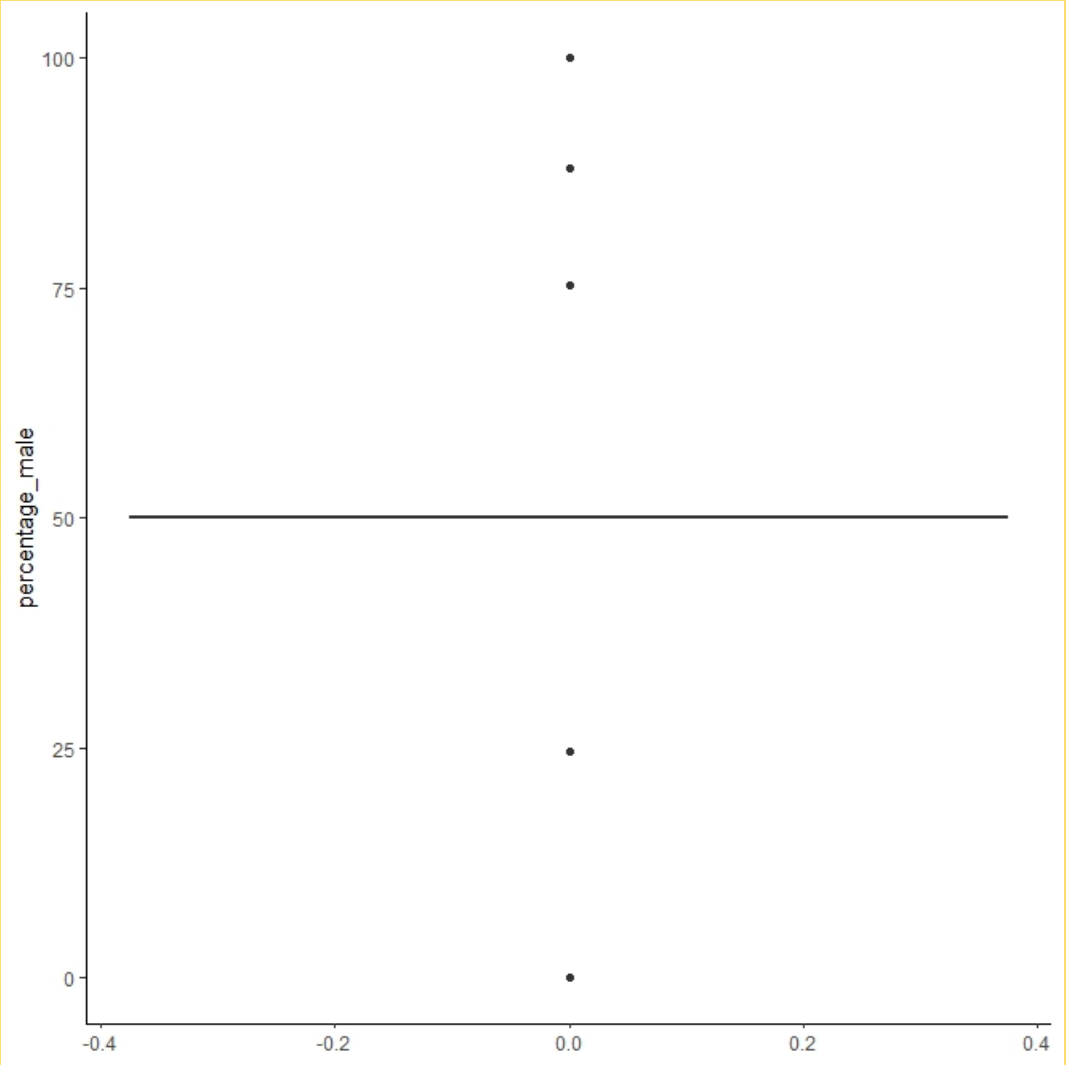
Check for potential outliers





Check for outliers





‘Variable not variable', let's not consider these parameters.

The key stages of the linear model

Definition of the question

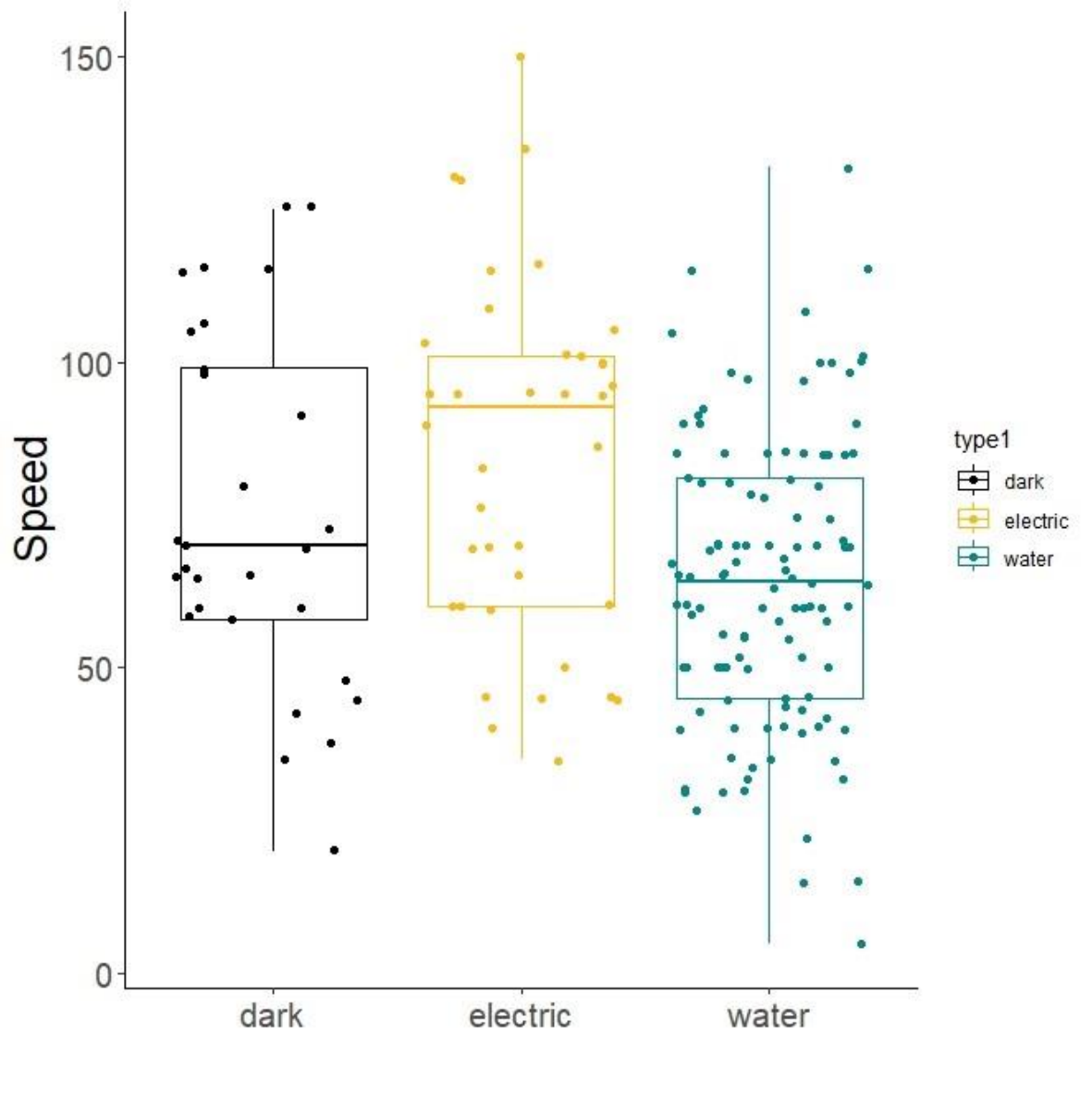
Exploration

Qualitative variables ← Sampling effort

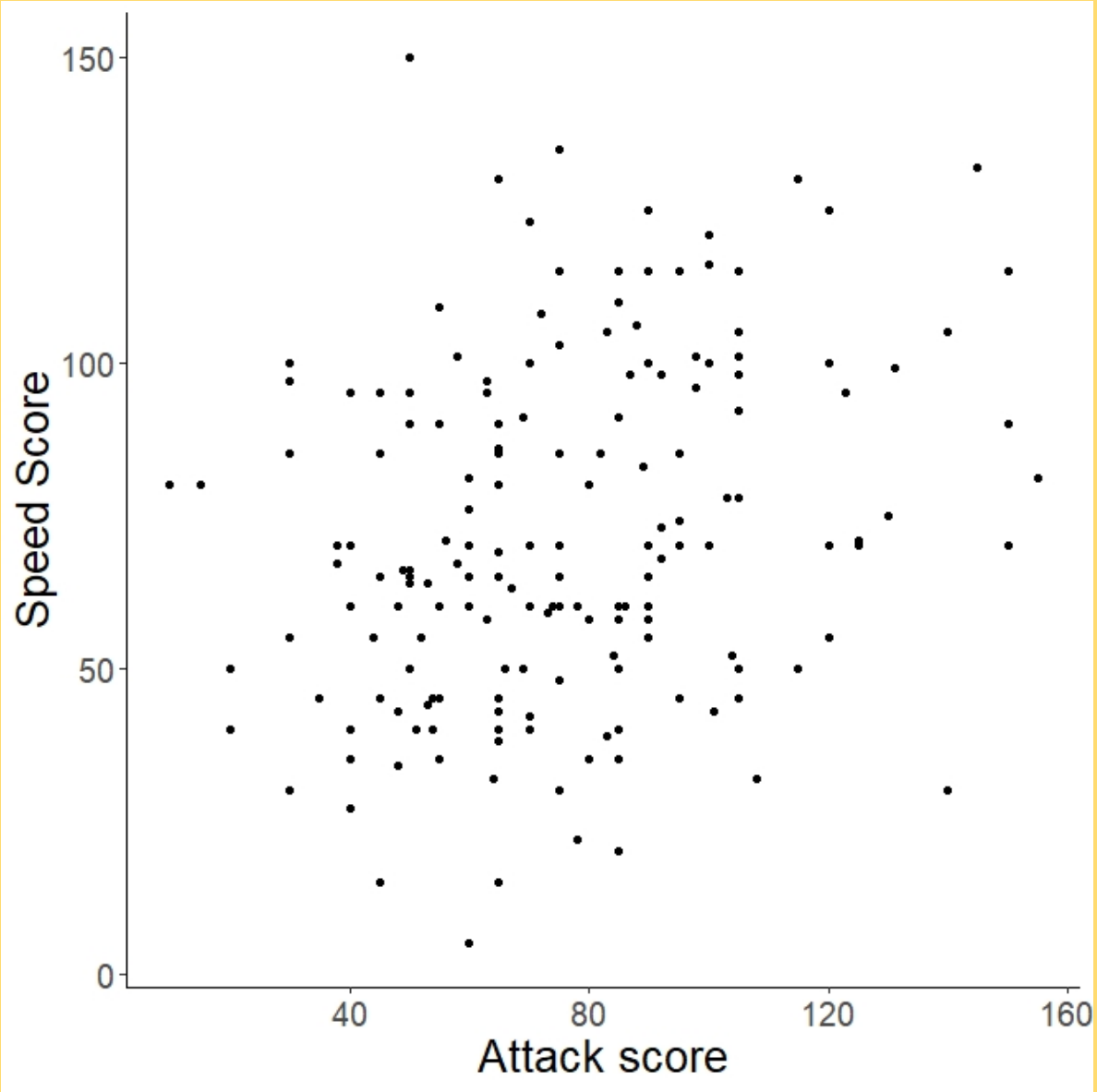
Quantitative variables ← outliers

relationships

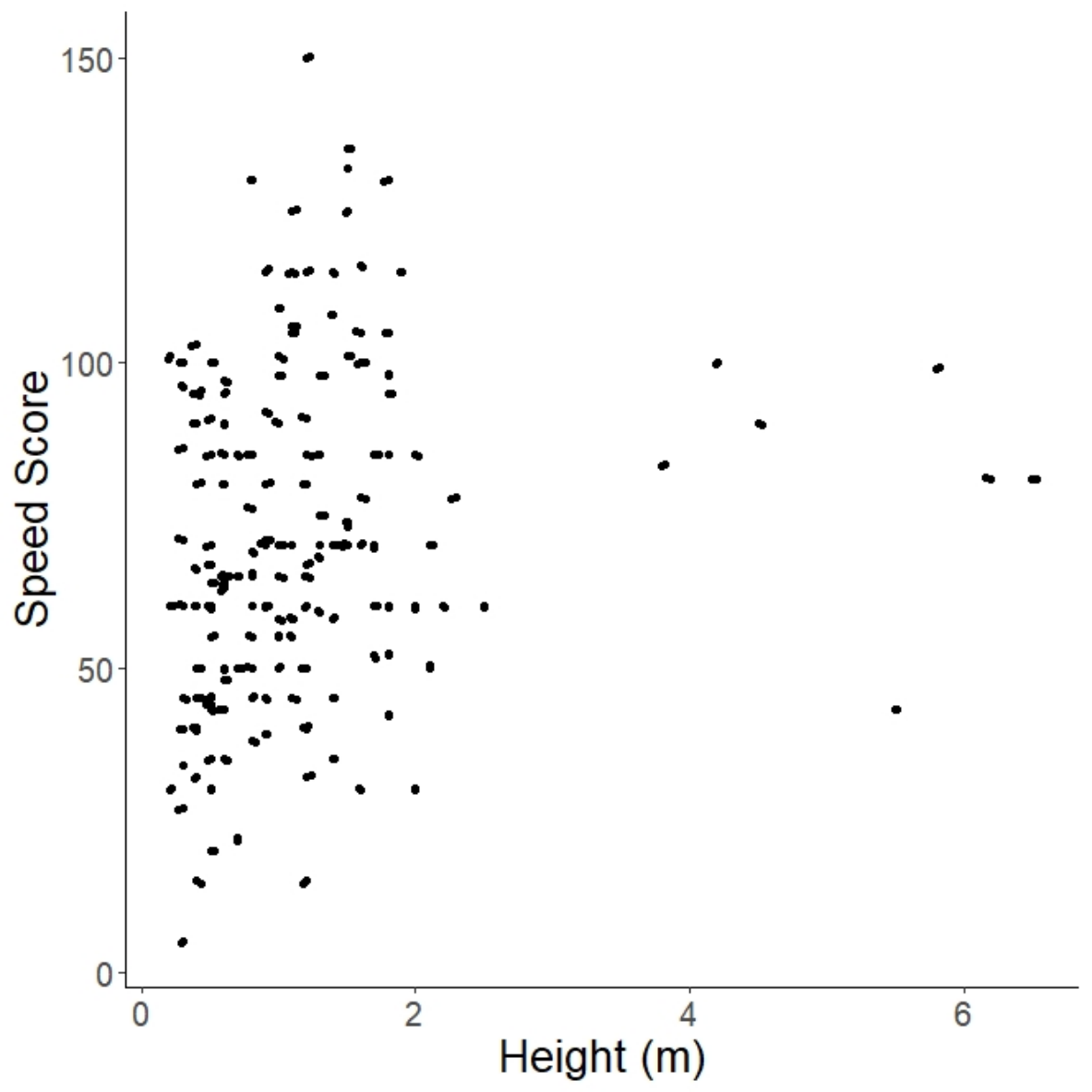
Speed as a function of Pokemon type



Speed as a function of attack score



Speed as a function of height



The key stages of the linear model

Definition of the question

Exploration

Model construction

Model construction

.

as a function of

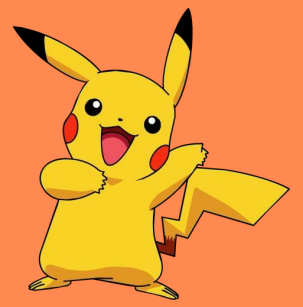
and

Response variable ~ Explanatory variable 1 + Explanatory variable 2 + ...

what I want to study

potential parameters
that I have considered

Speed ~ type + attack score + height



On R

response variable

‘as a function of’

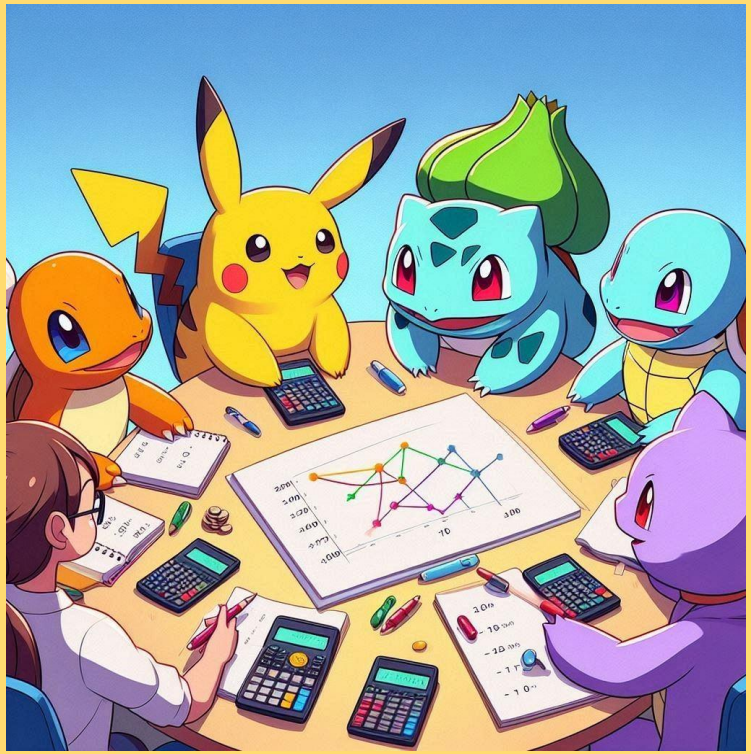
```
mod=lm(speed~height_m+attack+type1,
       data4)
```

calculations from the linear model stored in an object called ‘mod’

explanatory variables

interpretation

anova table → What are the significant effects?



On R

Response variable

```
> Anova(mod)
Anova Table (Type II tests)

Response: speed
          Sum Sq  Df F value    Pr(>F)
height_m    1269   1  2.0735  0.151663
attack      5169   1  8.4485  0.004126 **
type1      13777   2 11.2585 2.514e-05 ***
Residuals 107072 175
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

What can we conclude about significance ?

Explanatory variables

interpretation

anova table → What are the significant effects?

summary → Does the model explain a lot of variation?

About R

Does the model explain a lot of variation?

```
> #summary table
> summary(mod)

Call:
lm(formula = speed ~ height_m + attack + type1, data = data4)

Residuals:
    Min      1Q  Median      3Q     Max 
-53.413 -16.840  -1.282  17.503  69.055 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    52.0703     7.4634   6.977 5.98e-11 ***
height_m        3.0646     2.1282   1.440  0.15166    
attack          0.2195     0.0755   2.907  0.00413 ** 
type1electric  14.2247     6.2089   2.291  0.02315 *  
type1water     -7.7442     5.2466  -1.476  0.14173    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.74 on 175 degrees of freedom
Multiple R-squared:  0.1858,	Adjusted R-squared:  0.1672 
F-statistic: 9.986 on 4 and 175 DF,  p-value: 2.655e-07
```

On R

Does the model explain a lot of variation?

```
Residual standard error: 24.74 on 175 degrees of freedom
Multiple R-squared:  0.1858,    Adjusted R-squared:  0.1672
F-statistic: 9.986 on 4 and 175 DF,  p-value: 2.655e-07
```

Is the model significant? Is it of interest?

```
Residual standard error: 24.74 on 175 degrees of freedom
Multiple R-squared:  0.1858,    Adjusted R-squared:  0.1672
F-statistic: 9.986 on 4 and 175 DF,  p-value: 2.655e-07
```

Does the model, i.e. the explanatory variables considered, explain a large % of the variation in the response variable?

R^2 ~ % information

It's not dramatic to have a small R^2

interpretation

anova table → What are the significant effects?

summary → Does the model explain a lot of variation?

summary → Going further into significant effect

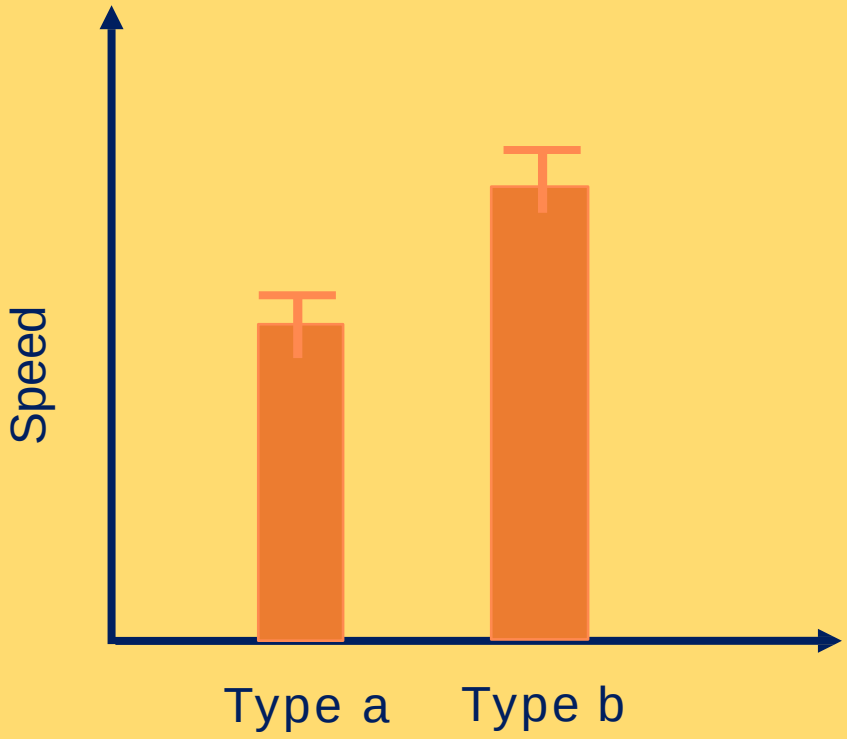
quali

quanti

Qualitative variable = Mean difference

H0 : there is no difference between the groups and equality between means

H1 : there is a difference between the means and groups

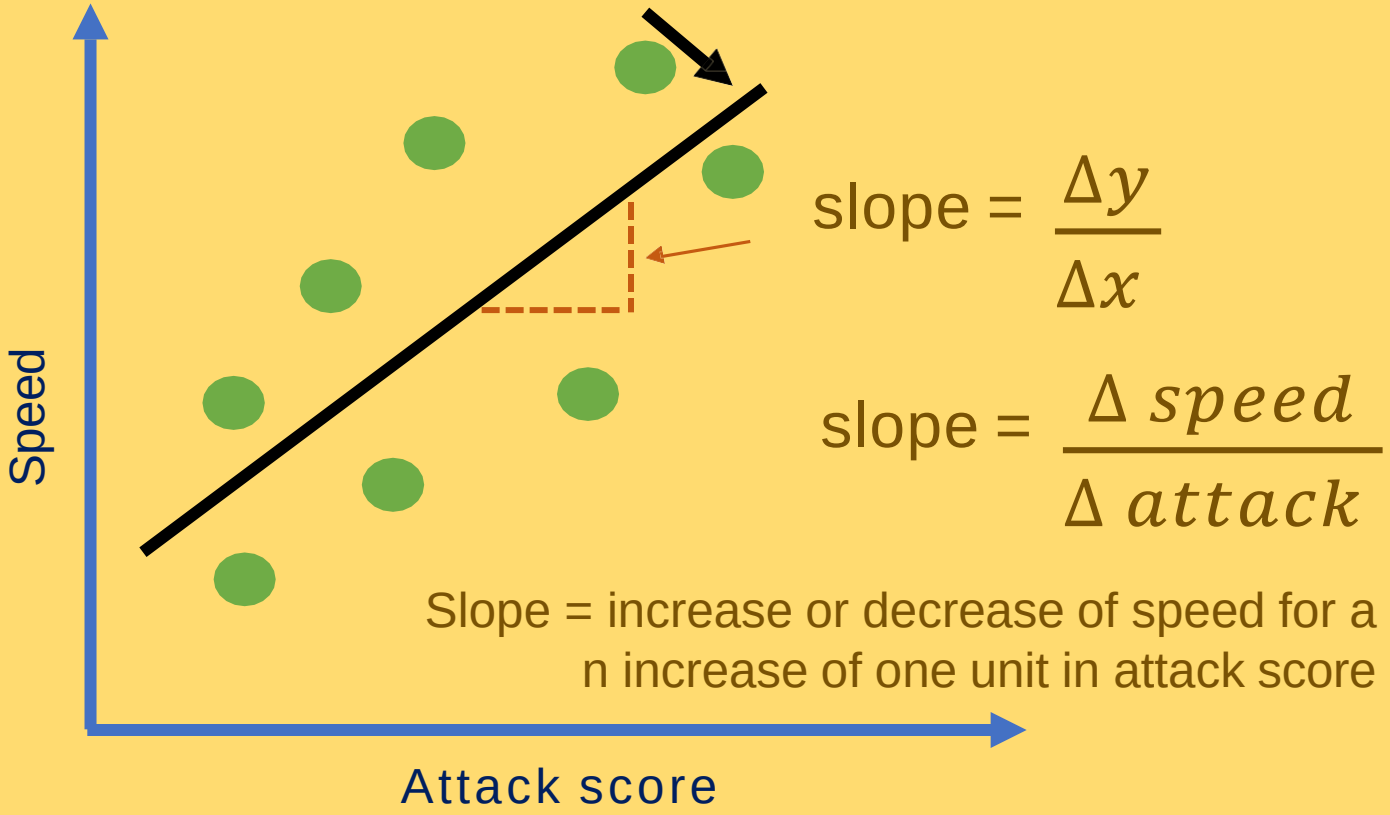


Quantitative variable = Slope

H0 : slope not different from 0

H1 : slope different from 0

Regression line whose slope corresponds to the coefficient



$$\text{slope} = \frac{\Delta y}{\Delta x}$$

$$\text{slope} = \frac{\Delta\ speed}{\Delta\ attack}$$

Slope = increase or decrease of speed for a n increase of one unit in attack score

Speed as a function of attack score

quanti

```
> #summary table
> summary(mod)

Call:
lm(formula = speed ~ height_m + attack + type1, data = data4)

Residuals:
    Min      1Q  Median      3Q     Max 
-53.413 -16.840  -1.282  17.503  69.055 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     52.0703     7.4634   6.977 5.98e-11 ***
height_m         3.0646     2.1282   1.440  0.15166    
attack           0.2195     0.0755   2.907  0.00413 ** 
type1electric   14.2247     6.2089   2.291  0.02315 *  
type1water      -7.7442     5.2466  -1.476  0.14173    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.74 on 175 degrees of freedom
Multiple R-squared:  0.1858,	Adjusted R-squared:  0.1672 
F-statistic: 9.986 on 4 and 175 DF,  p-value: 2.655e-07
```

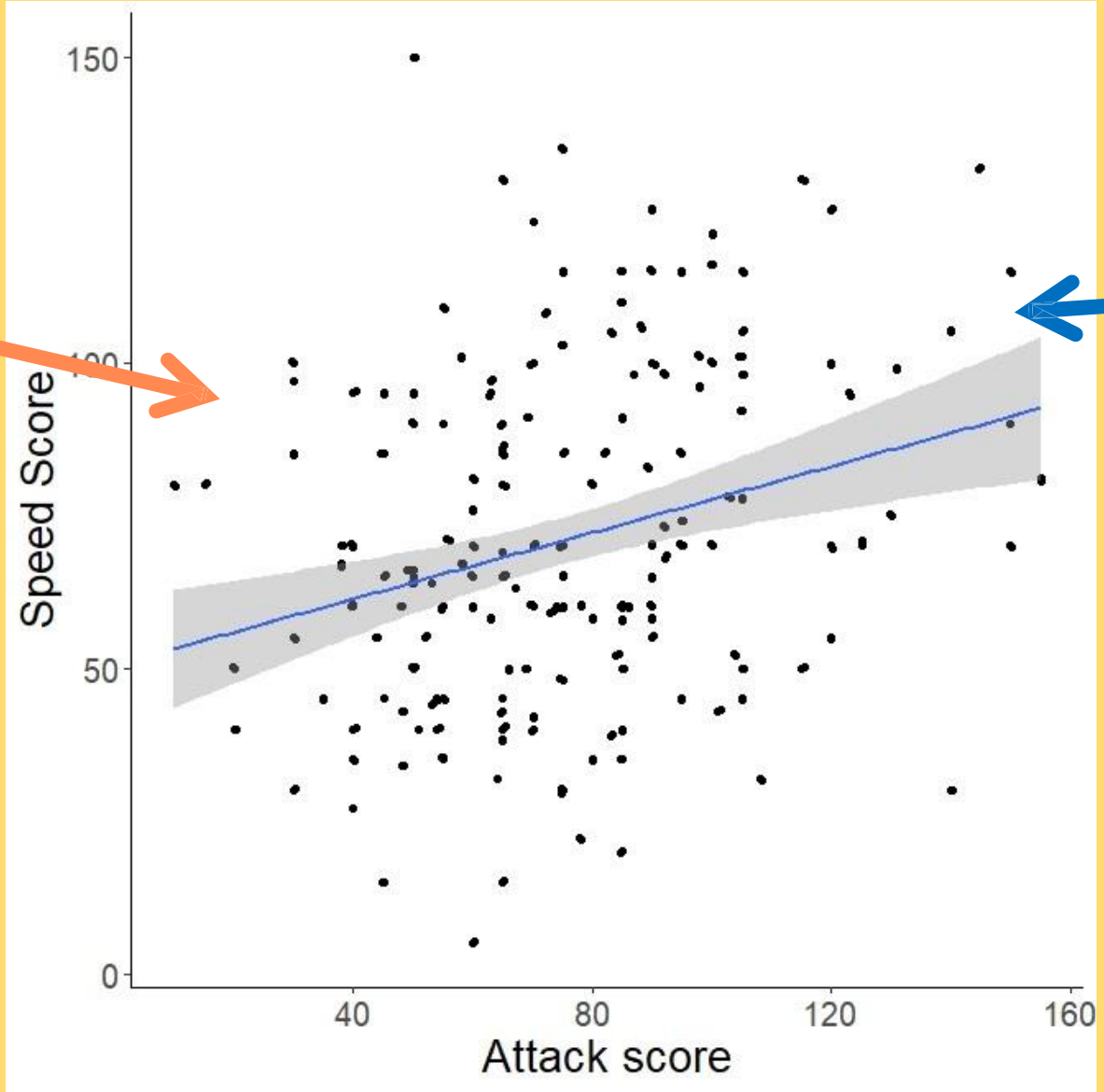
Step 1: significance

Step 2: slope coefficient valence

Step 3: slope 'effect size' interpretation

for an increase of one attack unit, there is an increase of 0.22 for the speed score

Speed as a function of attack score



Speed as a function of type

quanti

```
> #summary table
> summary(mod)

Call:
lm(formula = speed ~ height_m + attack + type1, data = data4)

Residuals:
    Min      1Q  Median      3Q     Max 
-53.413 -16.840  -1.282  17.503  69.055 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     52.0703     7.4634   6.977 5.98e-11 ***
height_m         3.0646     2.1282   1.440  0.15166    
attack           0.2195     0.0755   2.907  0.00413 ** 
type1electric   14.2247     6.2089   2.291  0.02315 *  
type1water      -7.7442     5.2466  -1.476  0.14173    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.74 on 175 degrees of freedom
Multiple R-squared:  0.1858,	Adjusted R-squared:  0.1672 
F-statistic: 9.986 on 4 and 175 DF,  p-value: 2.655e-07
```

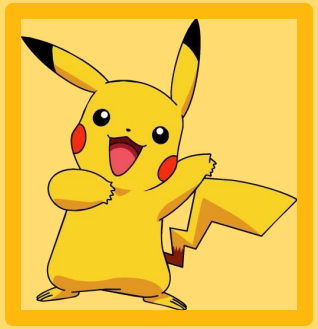
Step 1: significativity

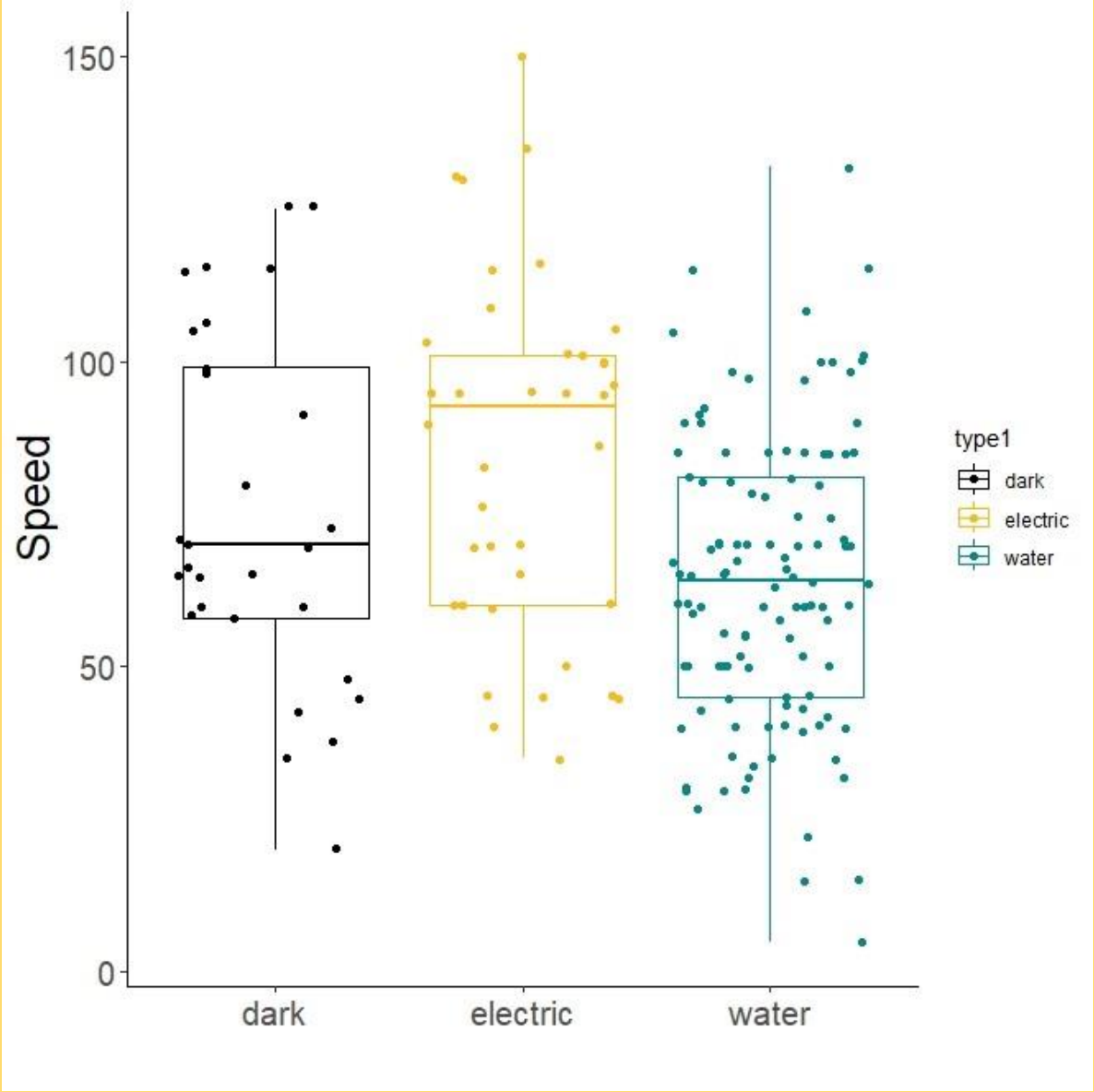
For some comparisons only

Step 2: mean difference signs

Step 3: mean absolute difference

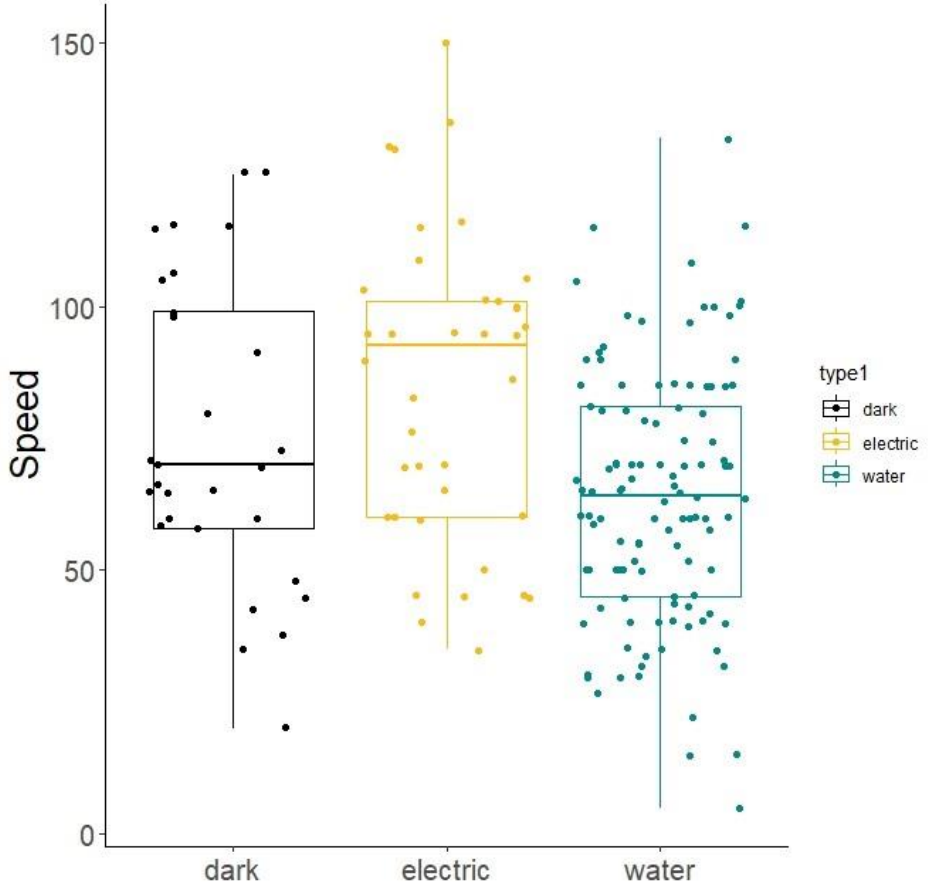
Speed as a function of Pokemon type





Post-hoc tests

```
> # post hocs ----------------------------------
> tcm=emmeans(mod,pairwise~type1)
> tcm
$emmeans
 type1    emmean   SE  df lower.CL upper.CL
 dark       72.0 4.67 175     62.8     81.3
 electric   86.3 4.03 175     78.3     94.2
 water      64.3 2.33 175     59.7     68.9

Confidence level used: 0.95

$contrasts
 contrast         estimate   SE  df t.ratio p.value
 dark - electric    -14.22 6.21 175  -2.291  0.0597
 dark - water         7.74 5.25 175   1.476  0.3049
 electric - water    21.97 4.65 175   4.723  <.0001

P value adjustment: tukey method for comparing a family of 3 estimates
```



P values corrections

The more you search, the more you find

Increase risk of type 1 errors with multiple testing.

=> p-value correction to keep the confidence level 5%

P-value corrections

- Several types :
  - **Bonferroni correction :** Adjusts p-values by multiplying them by the number of comparisons (N).
  - **Holm :** less severe than Bonferroni
  - **Tukey procedure :** after an ANOVA, the test considers every possible pair of means and computes a range of values within which the differences between group means must fall to be considered statistically significantly different.

The key stages of the linear model

Definition of the question

Exploration

Model construction

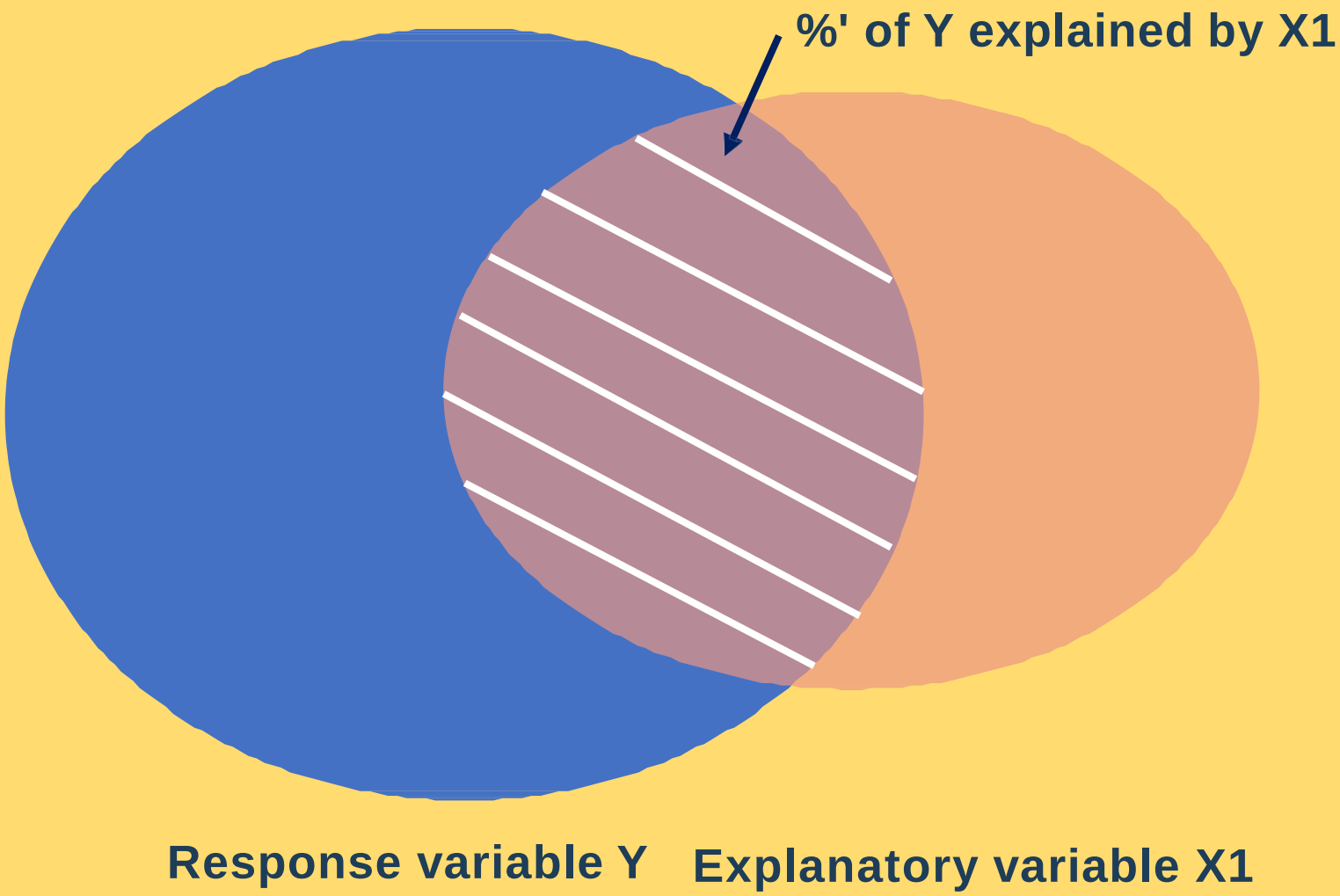
Checking application conditions

Checking application conditions

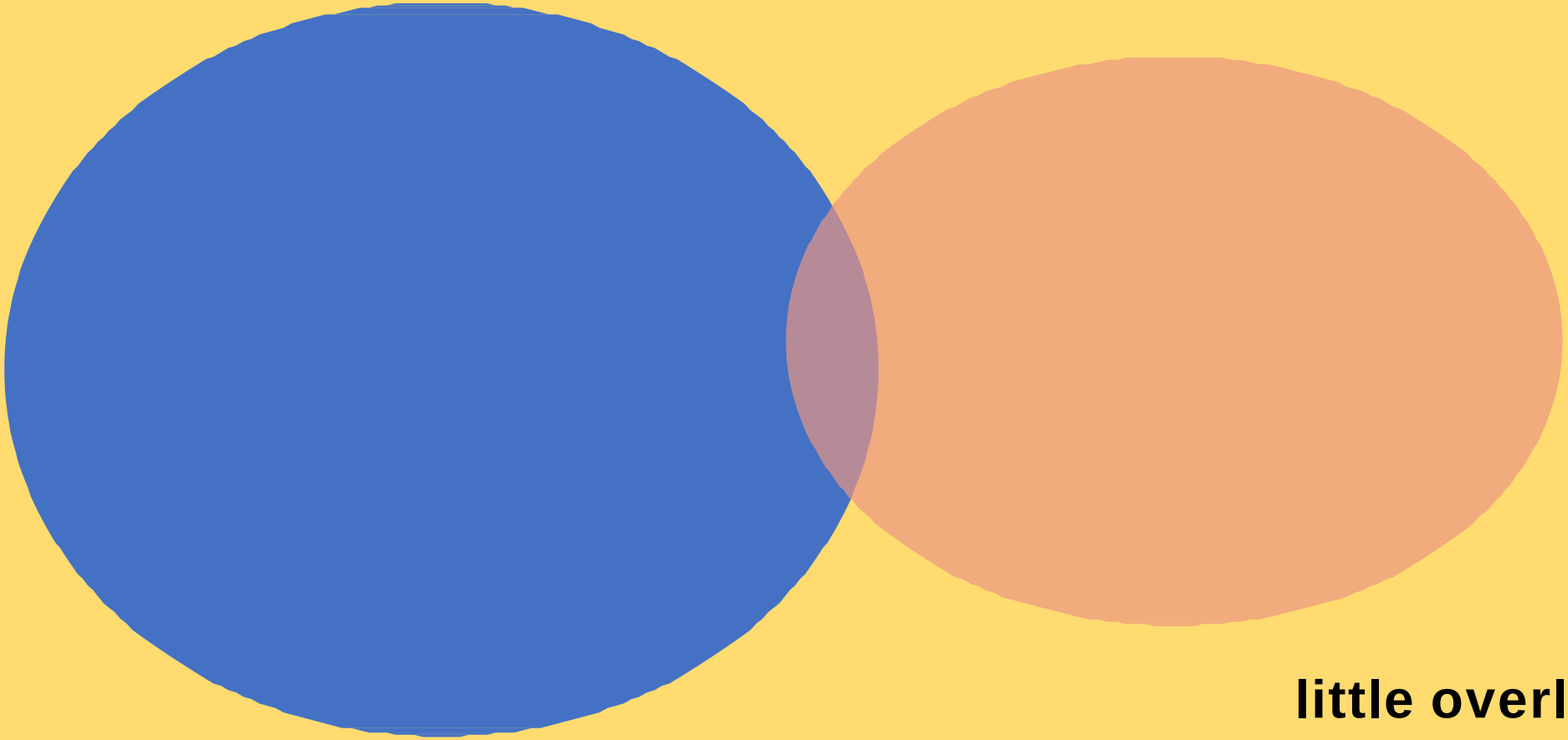
- Were we allowed to use the model?
- Can we trust the conclusions?

- **The conditions for applying a linear model :**
  - No colinearity between explanatory variables
  - Independance
  - Hypothesis on residuals :
    - Linearity
    - Normality
    - Homoscedasticity
    - Outliers

No collinearity expected between explanatory variables

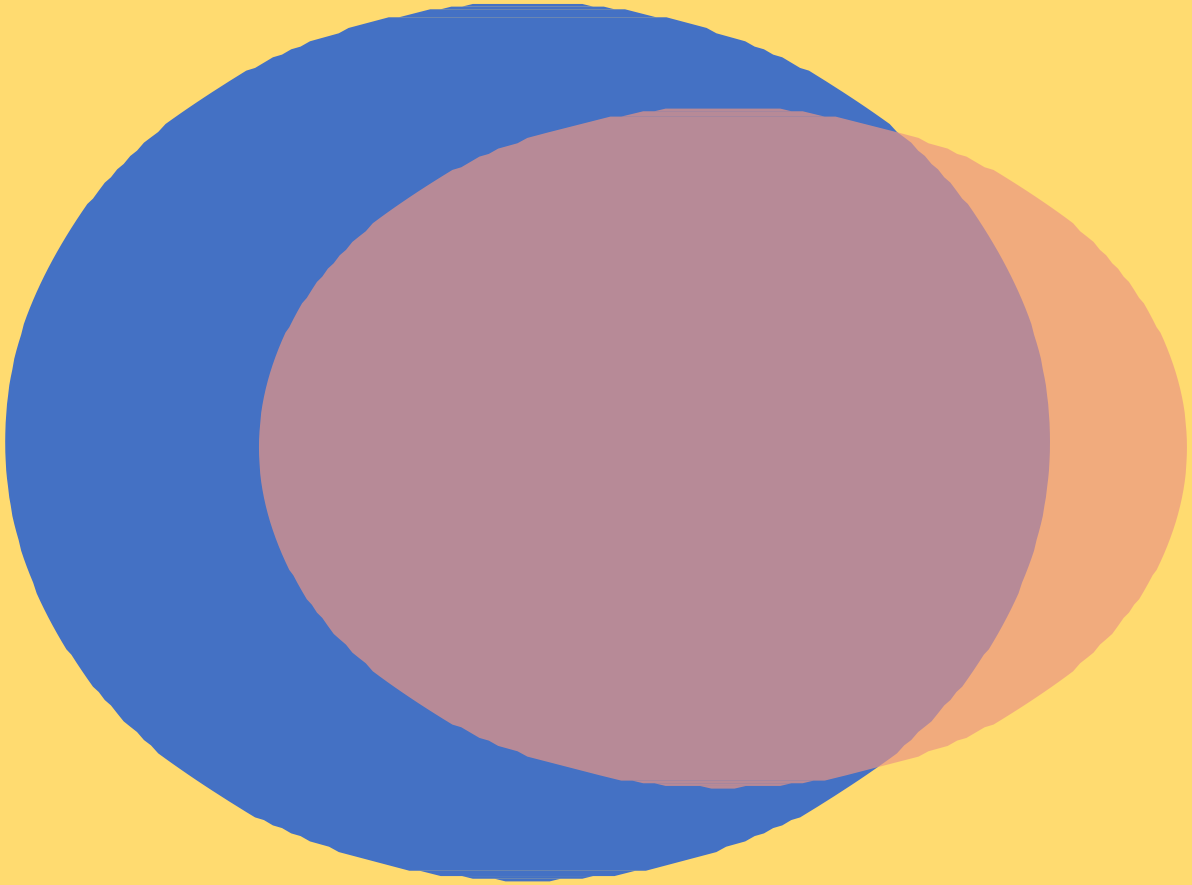


collinearity



little overlap
=> X1 explains little Y

collinearity

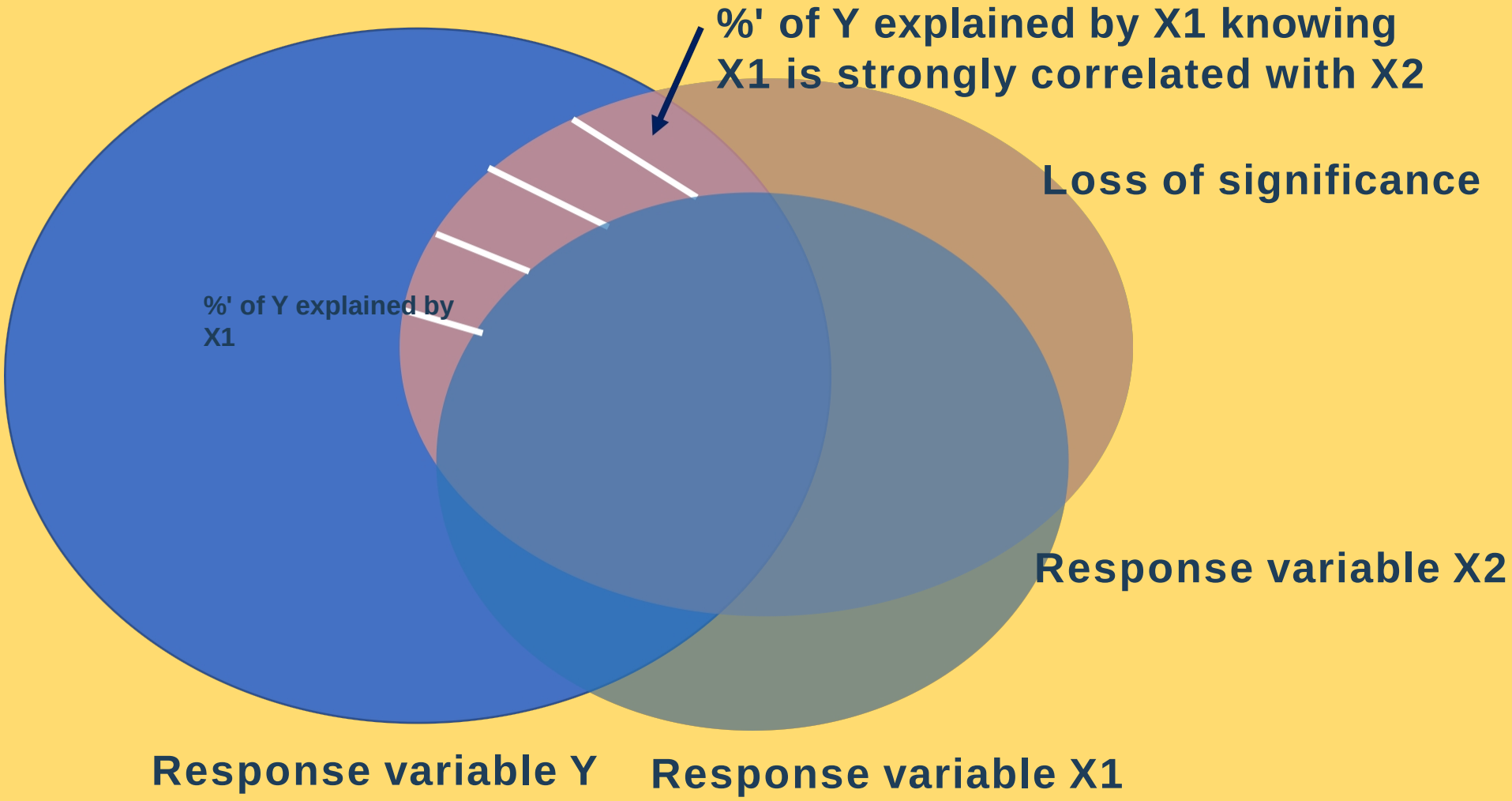


lots of overlap
=> X1 explains a lot about Y

collinearity

Increase in the value of standard errors

Inflated p-values

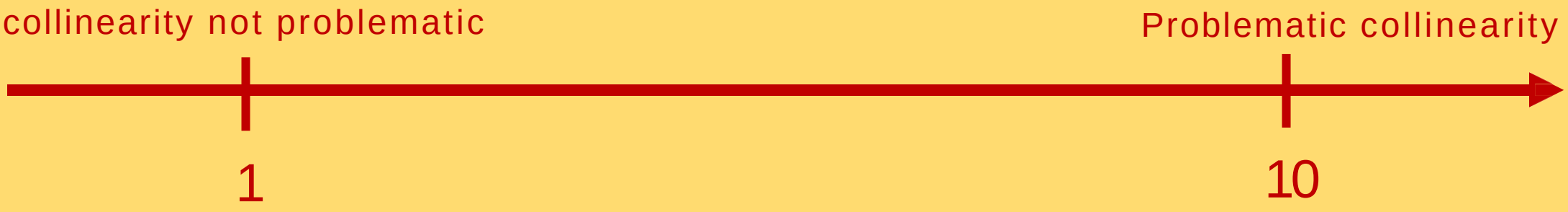


collinearity

If two people speak at the same time, you won't hear each other's message.

Check for potential collinearity

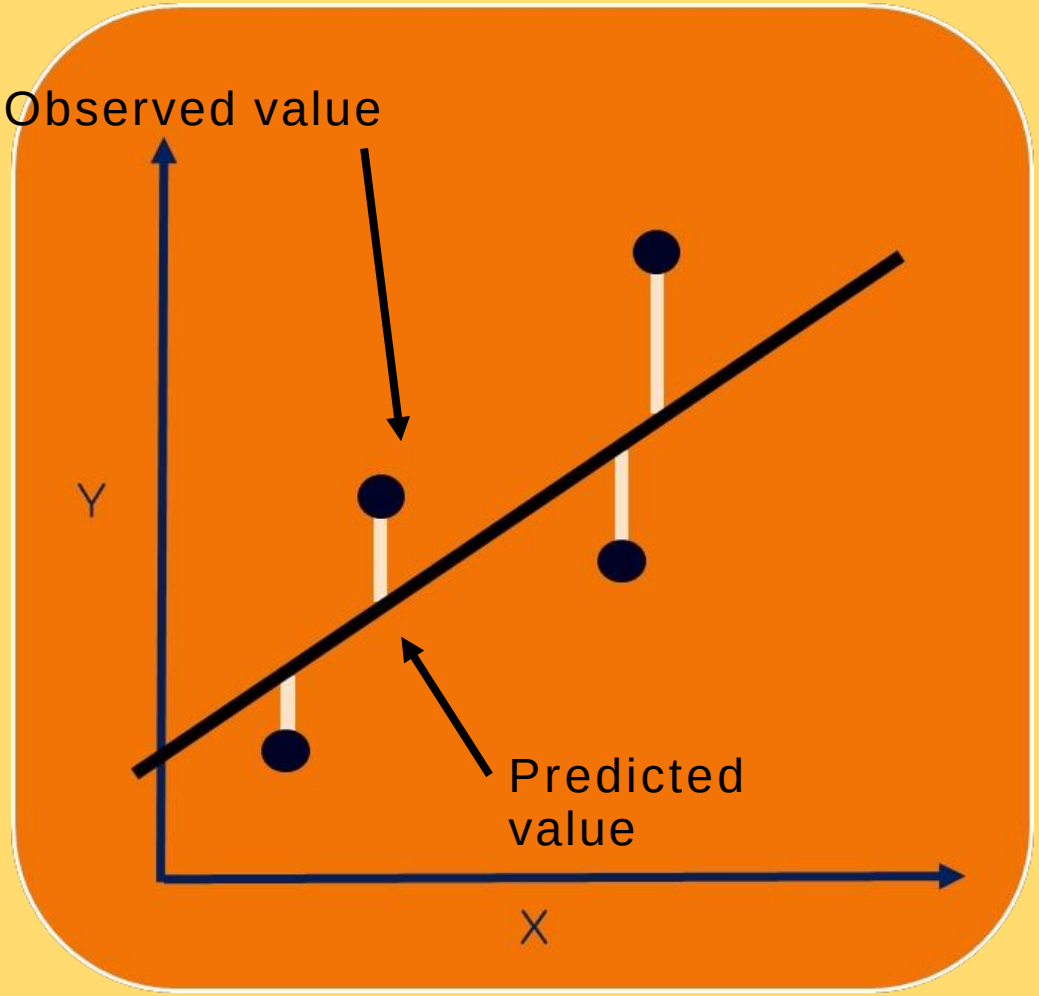
VIF (Variance Inflation Factor)

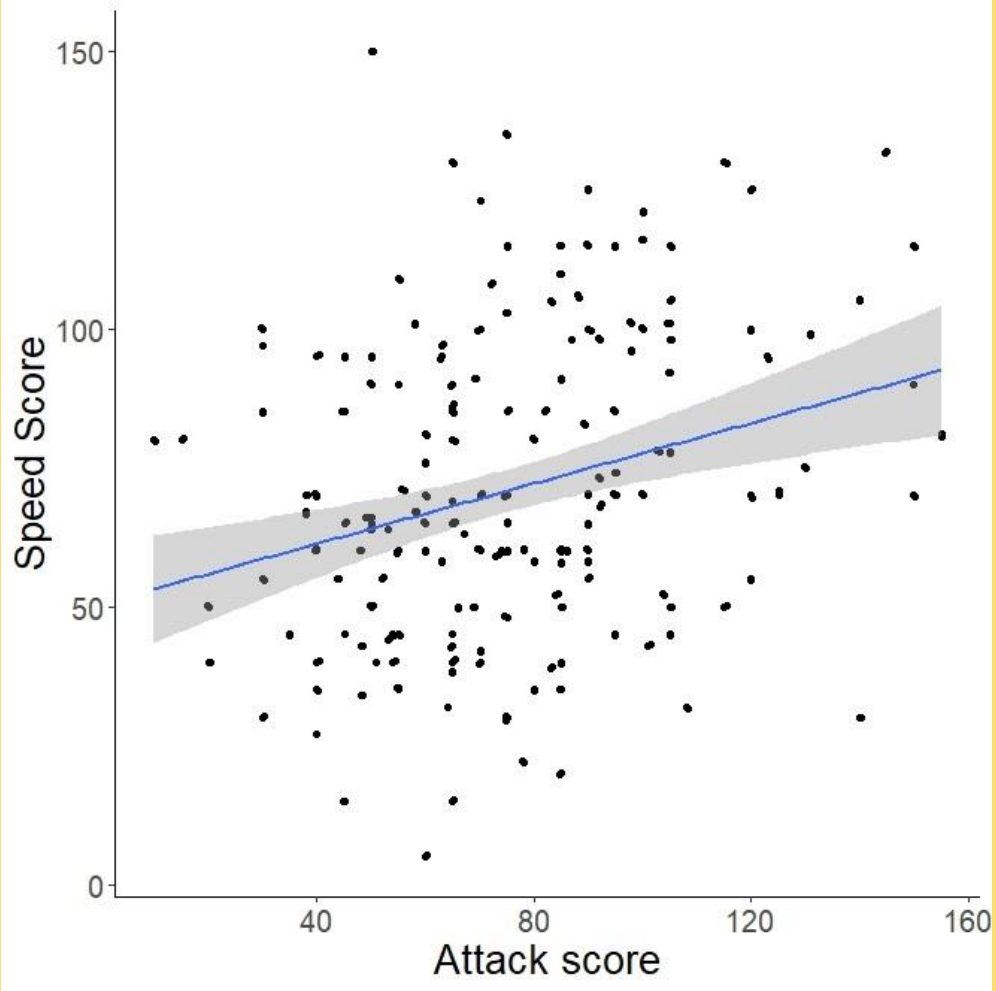



```
> # colinarity
> vif(mod)
             GVIF Df GVIF^(1/(2*Df))
height_m 1.286066  1        1.134049
attack   1.328377  1        1.152552
type1    1.048652  2        1.011947
```

VIF should be lower than 3

Model residuals = model errors





What's a mistake?

- ~ accident, sometimes the model crashes.
- It's when you don't on purpose.
- It's when it's random.

Residuals that indicate a healthy model?

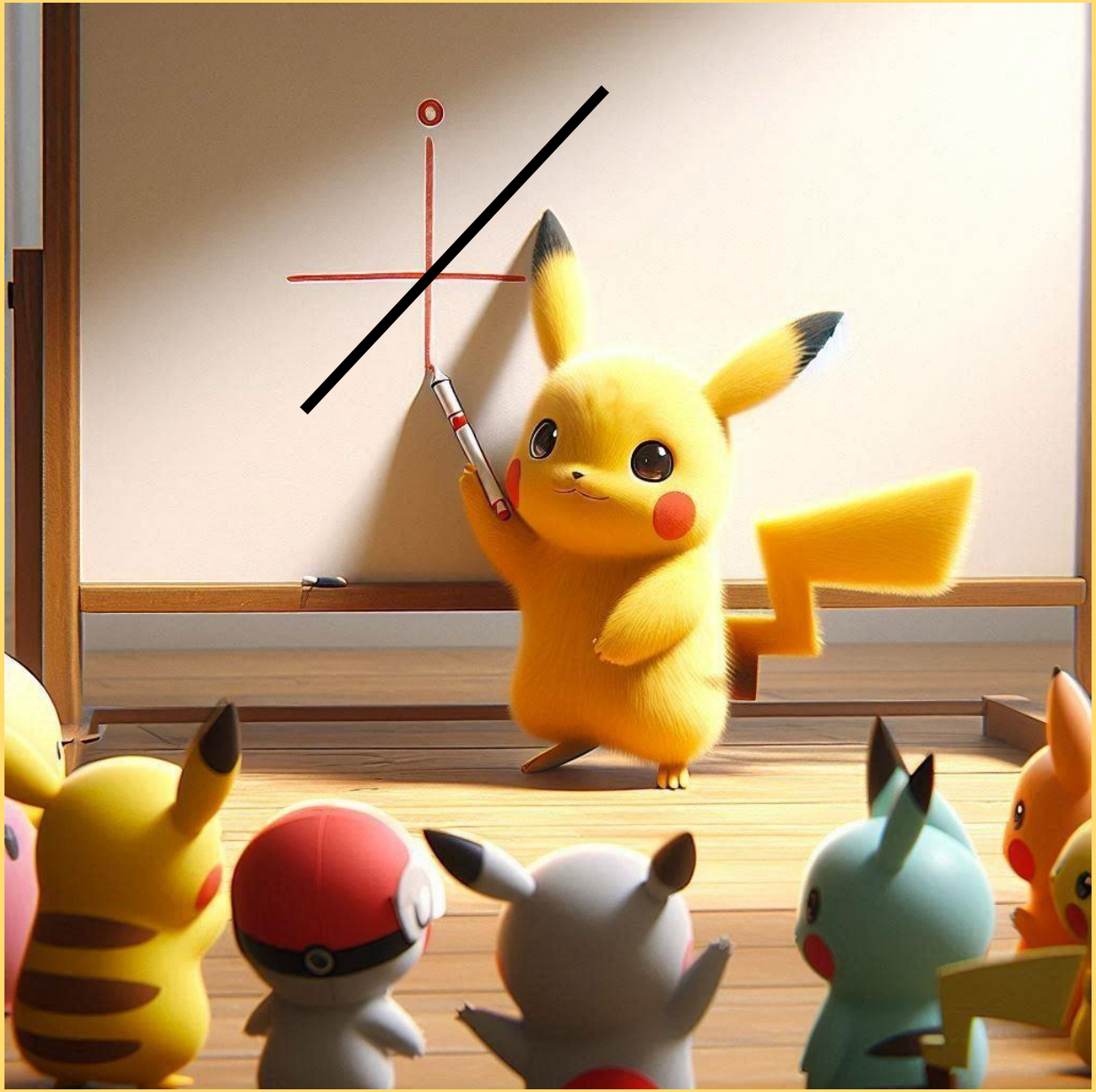
- distributed according to a Normal distribution with mean 0

- homogeneously distributed, i.e., with an even constant variance

- Linear relationships

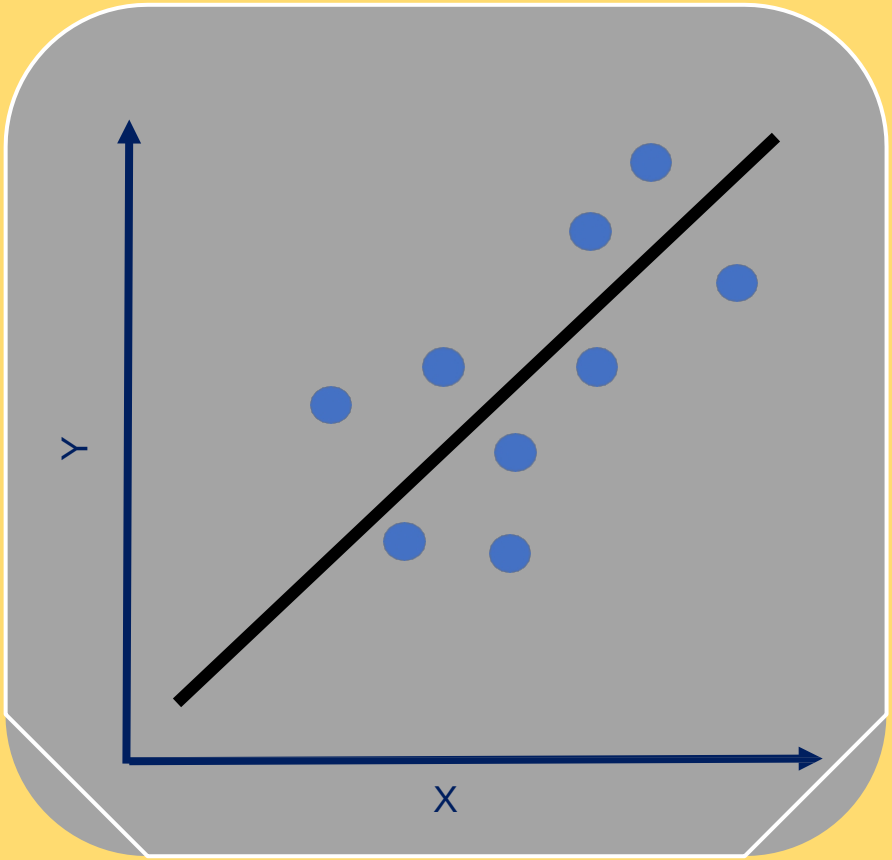
- No outlier

Linearity assumption: do we have a straight line?

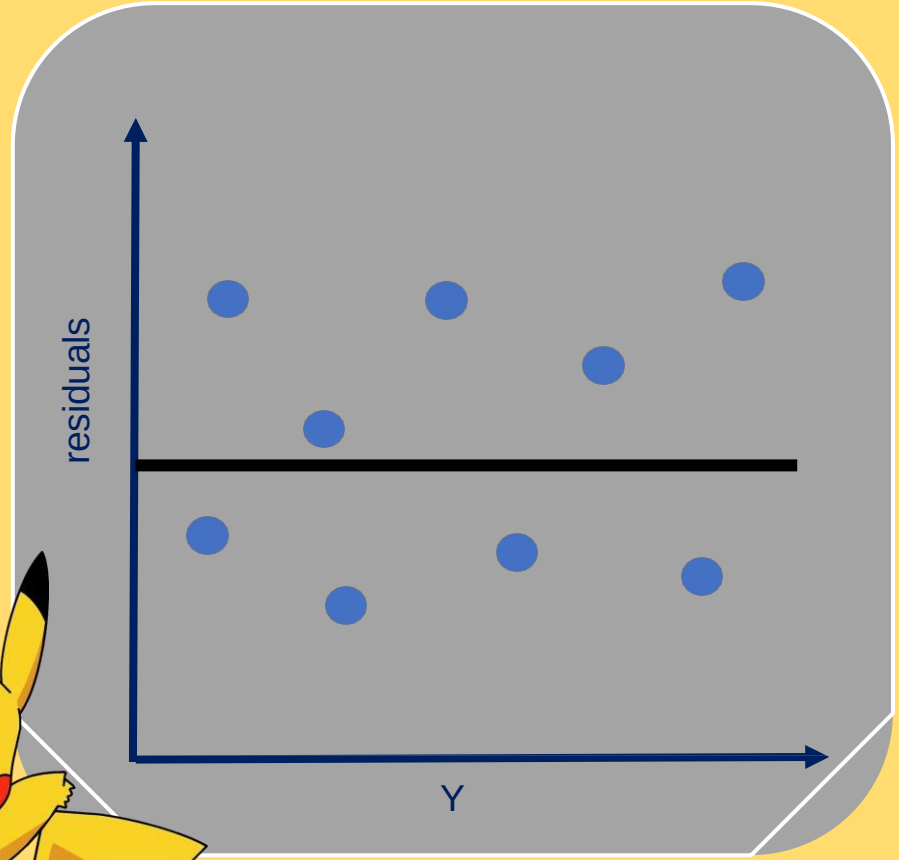


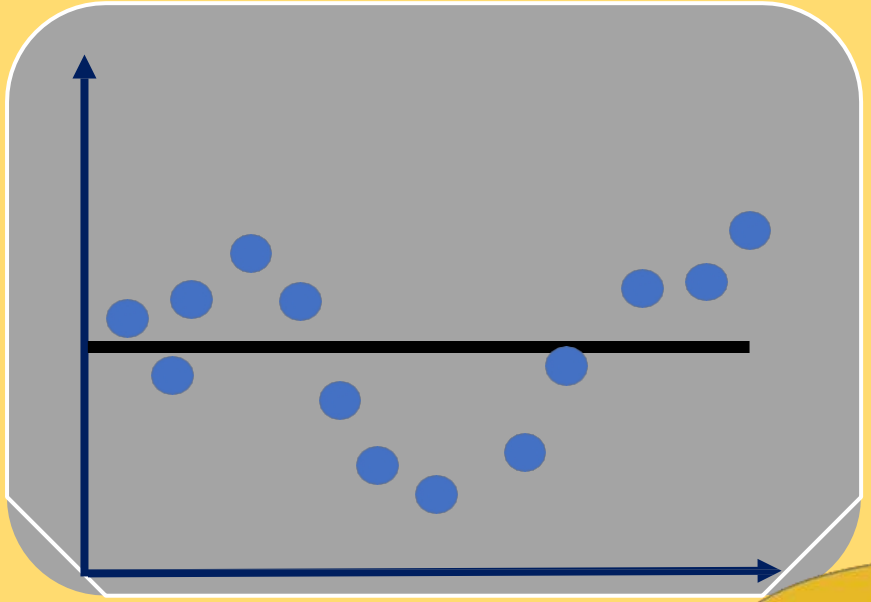
Linearity assumption: verification using a point cloud

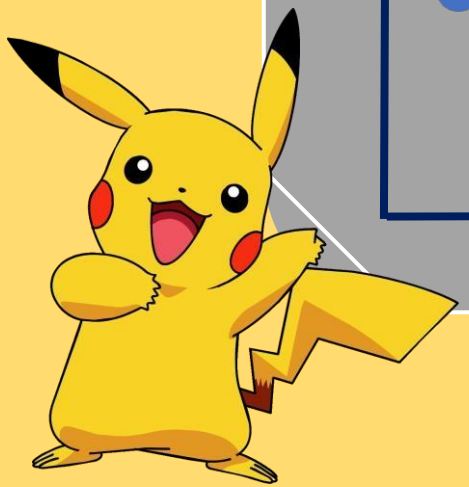
- Simple linear regression is used to examine the existence of a significant linear relationship between two numerical variables.
- Visual assessment: scatterplot.
- The shape of the scatterplot must approach a straight line.

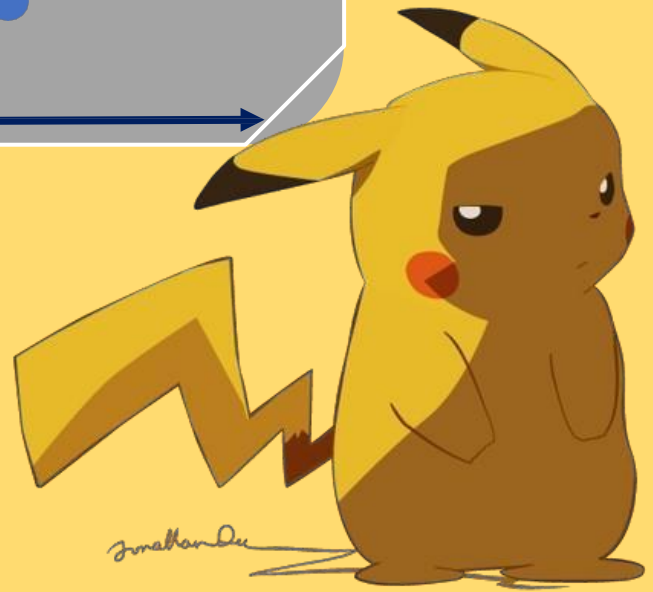


Linearity assumption: verification using residuals

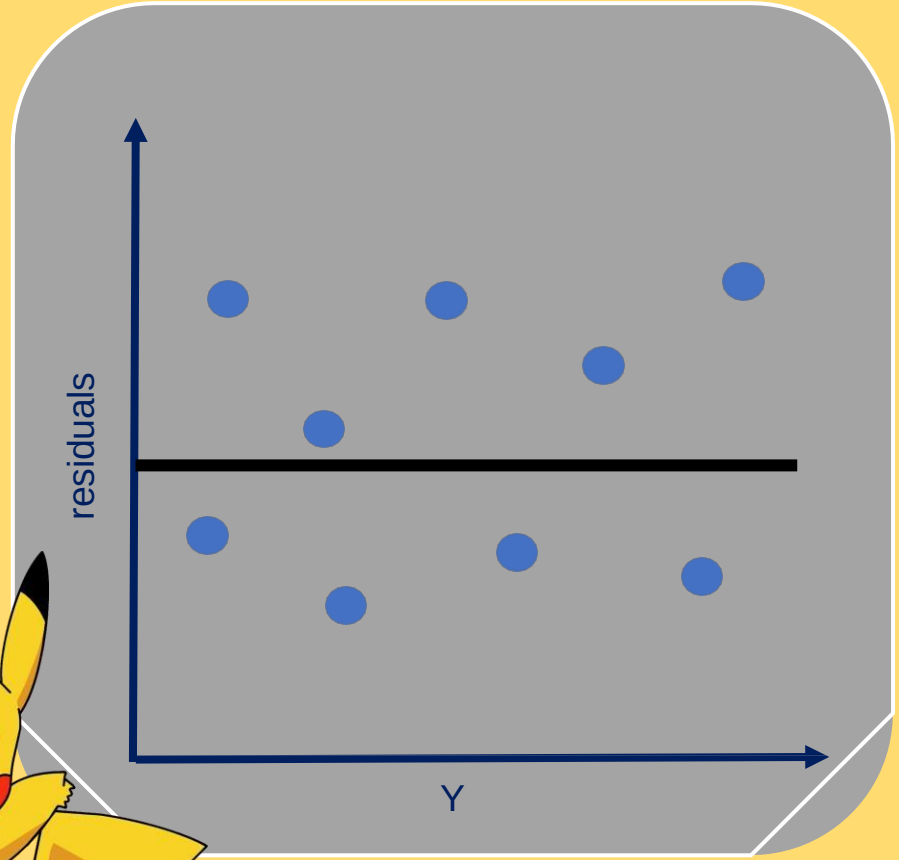


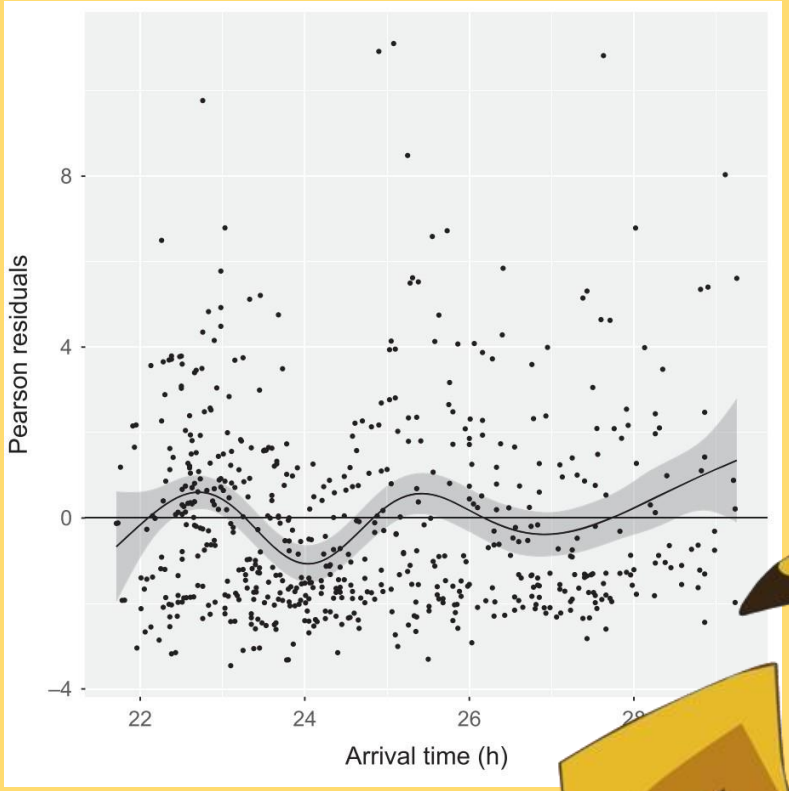




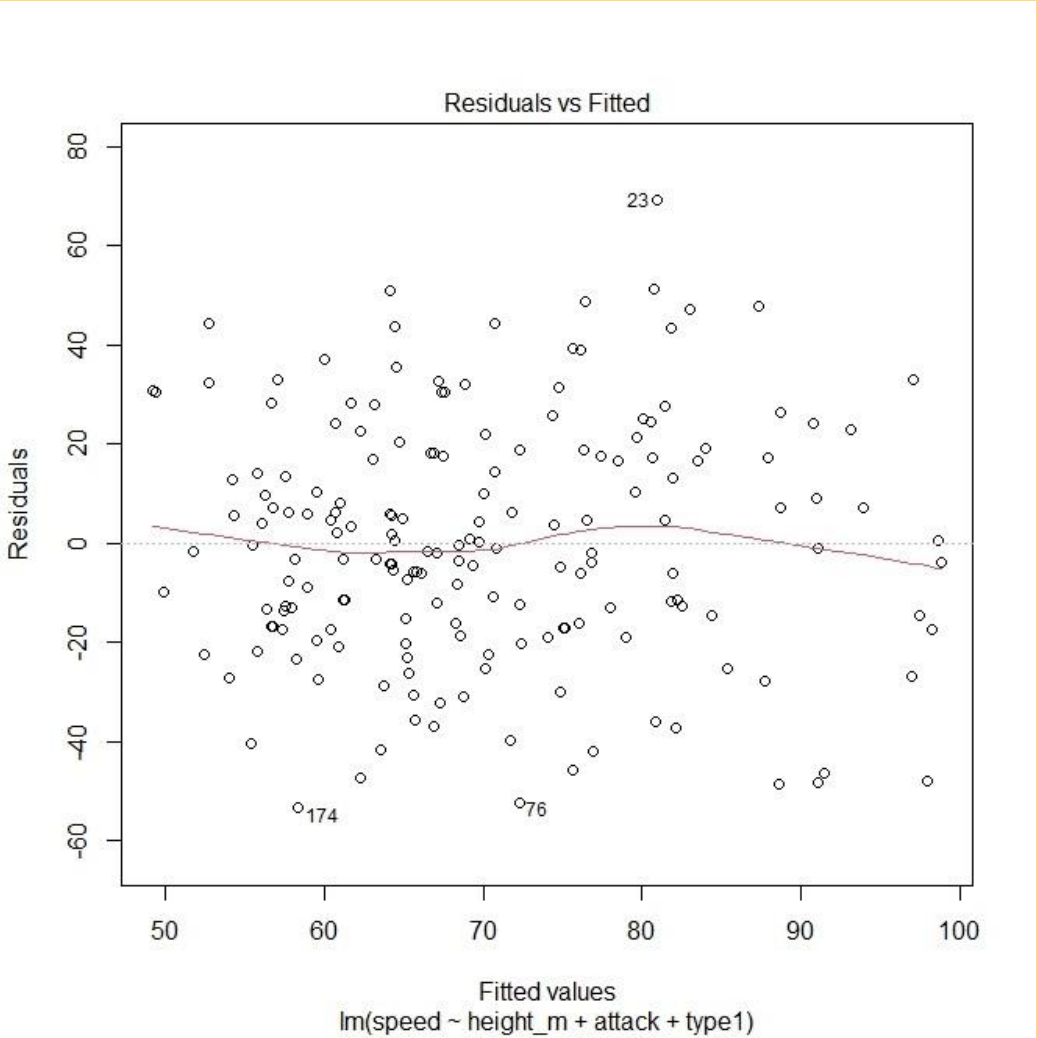


Linearity assumption: verification using residuals





Linearity assumption



Homogeneity/homoscedasticity of residuals

Model error variability must be constant

Errors should not be larger or smaller depending on variable values.

Avoid: *The higher the values of the response variable, the greater the errors (over- or under-estimation).*

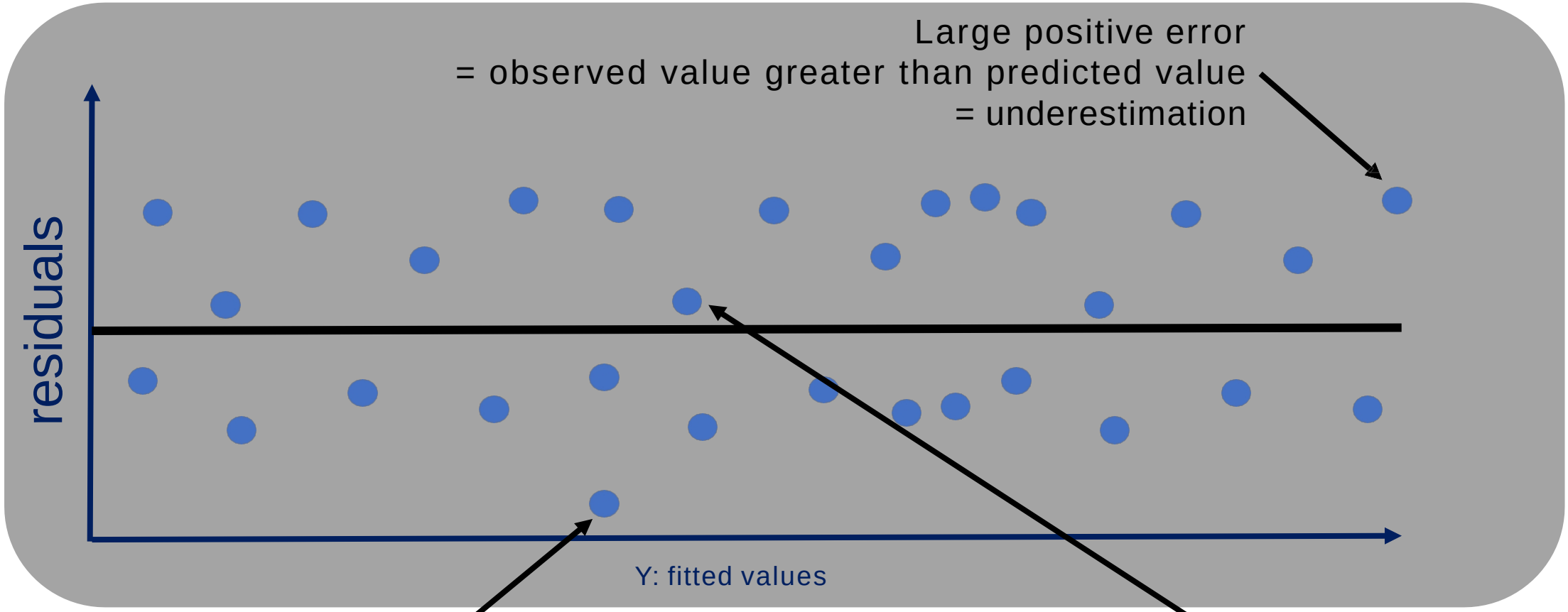
Homogeneity/homoscedasticity of residuals

How to spot it ?

Plot of *residuals vs fitted values*

Ok : The point cloud shows a constant variation along the horizontal axis (a tube shape is visualized).

Homogeneity/homoscedasticity of residuals (observed values - predicted values)



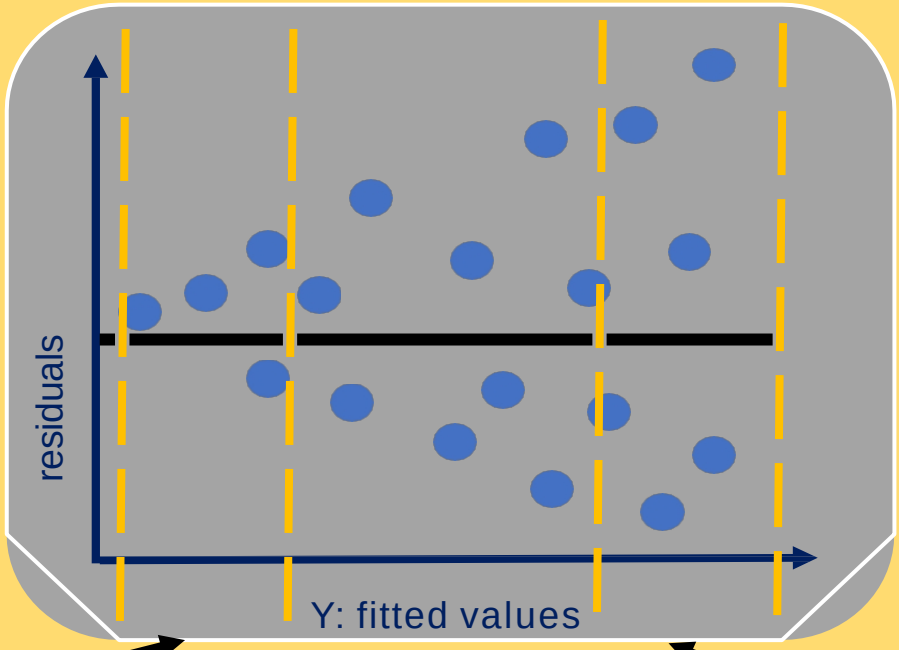
Large negative error

= observed value smaller than predicted value
= overestimation

small error in positive =

small underestimate

Homogeneity/homoscedasticity of residuals (observed values - predicted values)



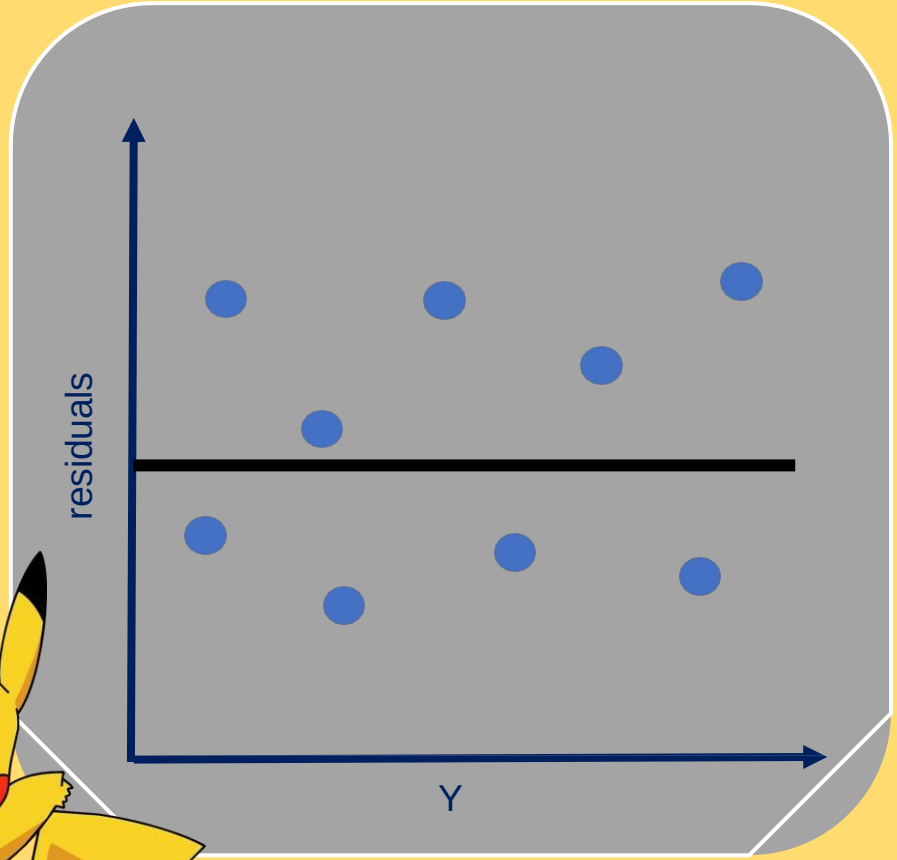
Small error variation when
Y values are small.

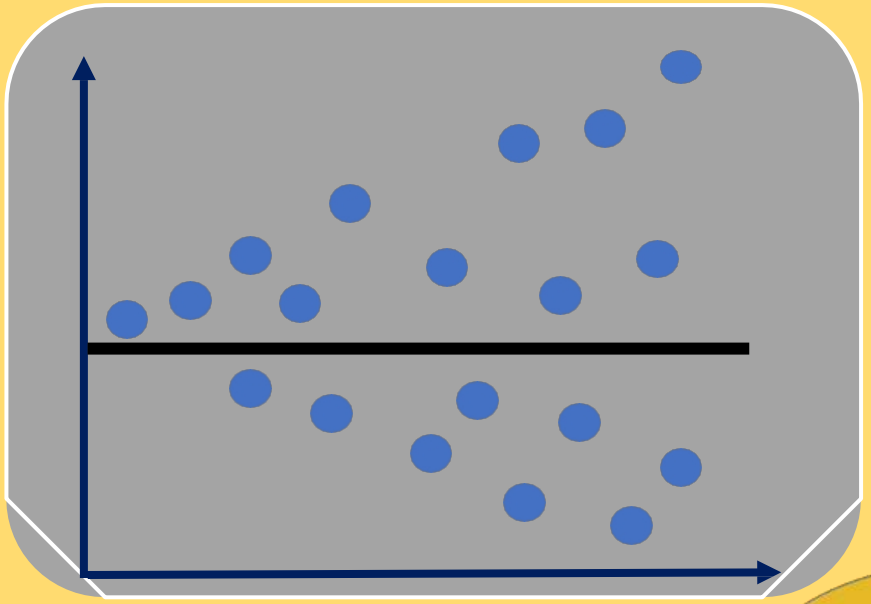
high error variation when
Y values are high.

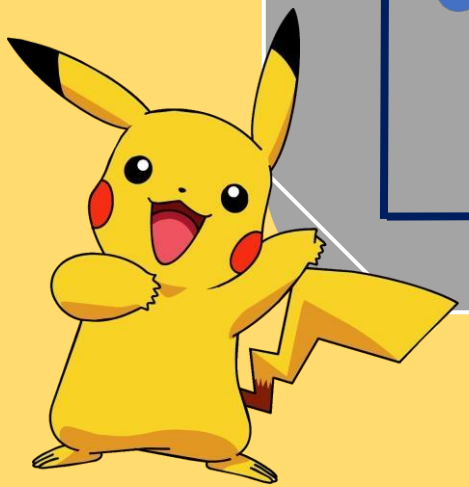
Wide range of errors when
Y values are large.
Big overestimates
and large underestimates.

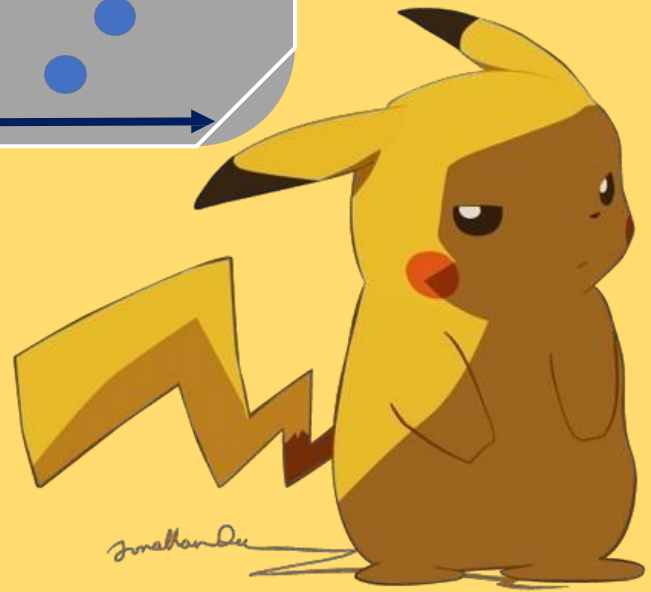
We make small overestimates
and small underestimates.

Homogeneity/homoscedasticity of residuals

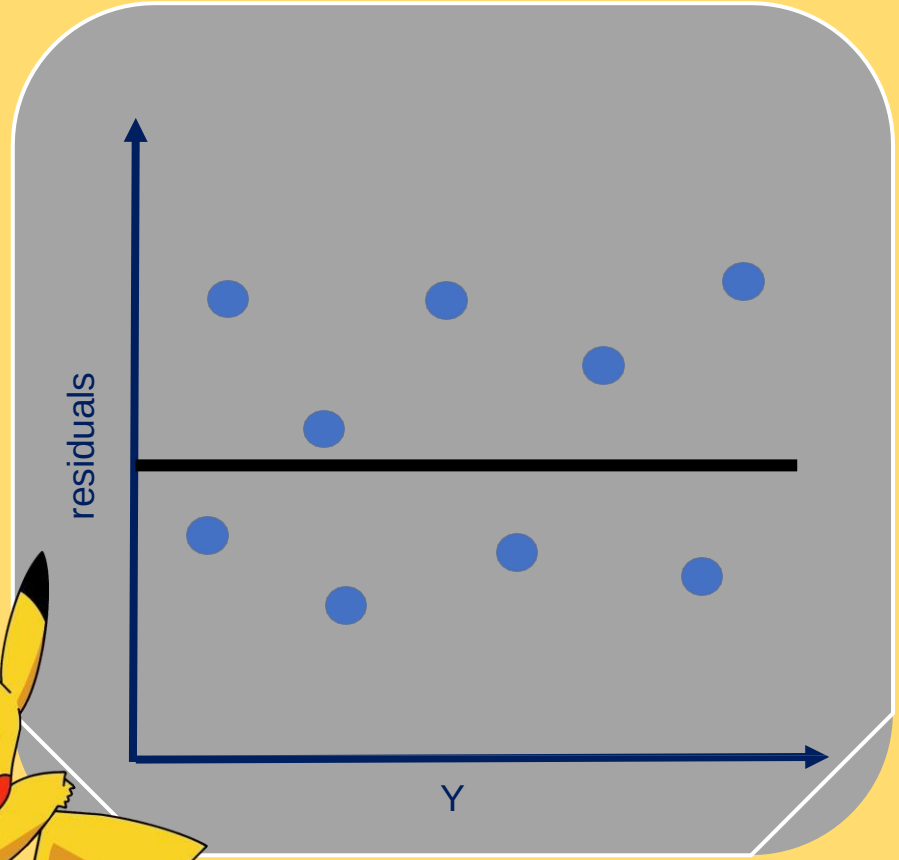


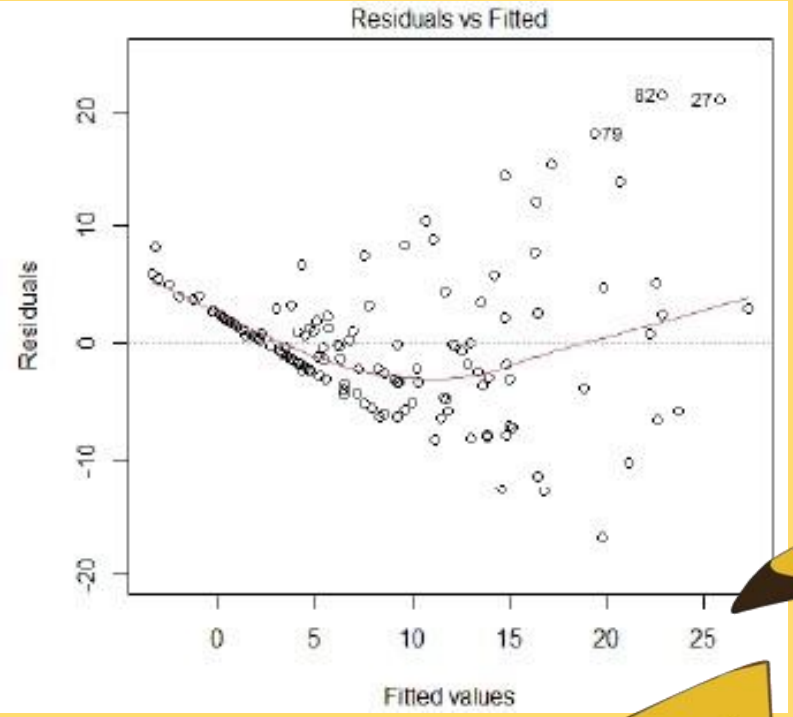


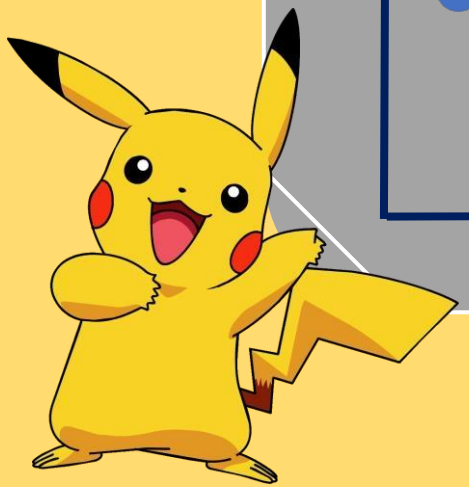


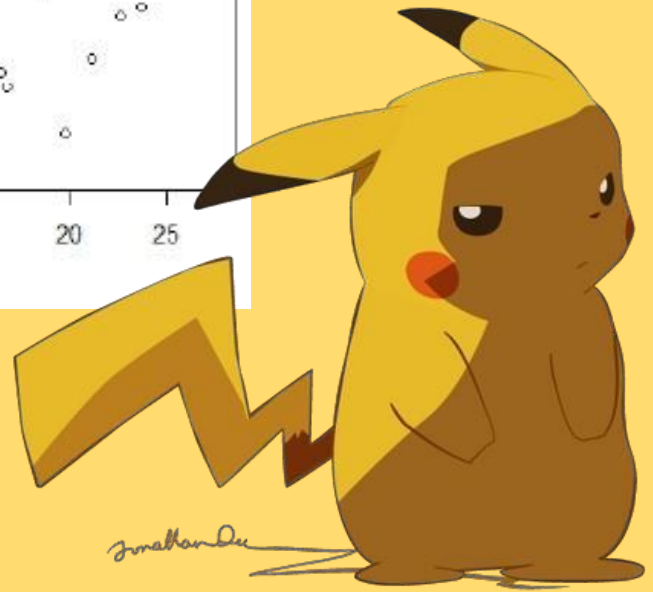


Homogeneity/homoscedasticity of residuals

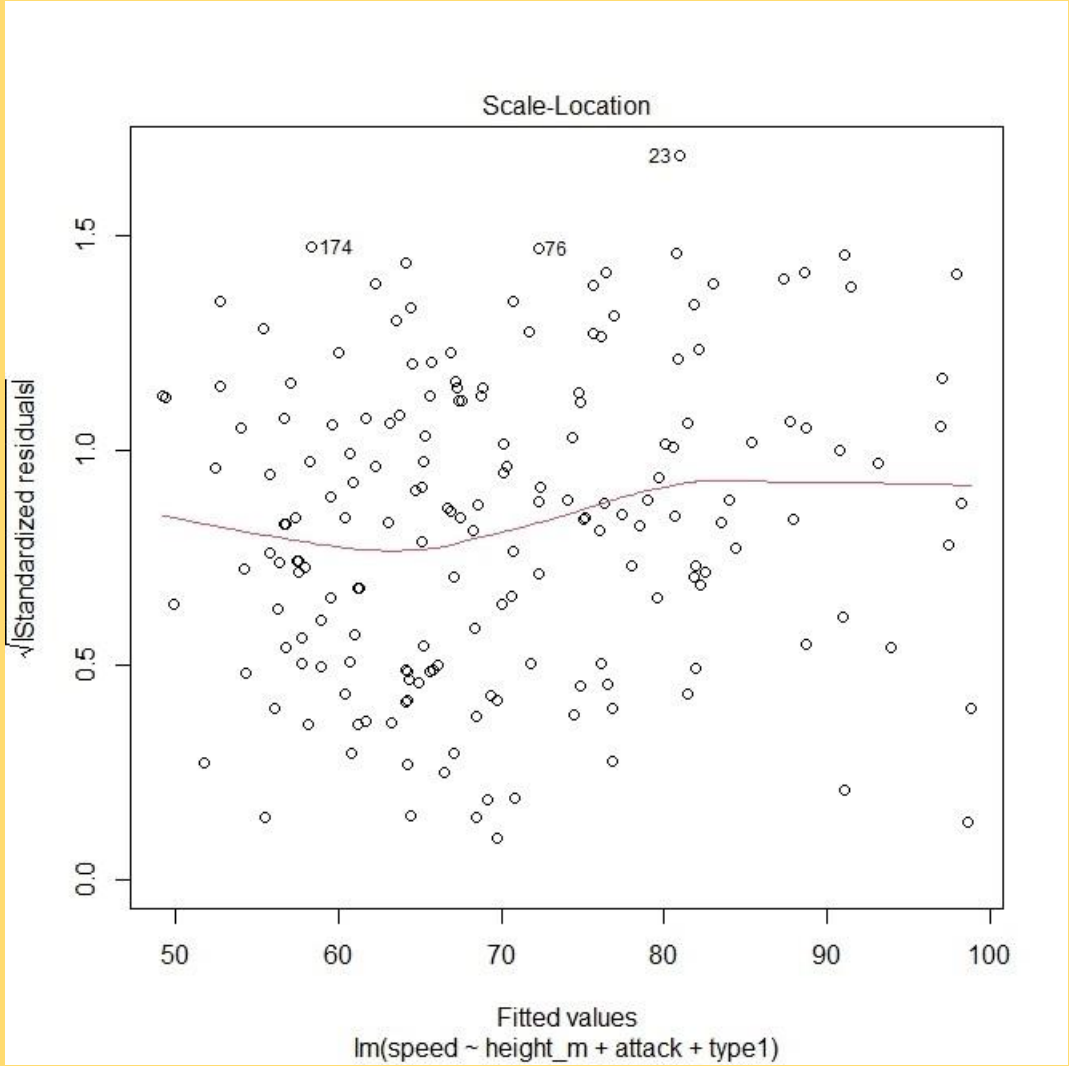






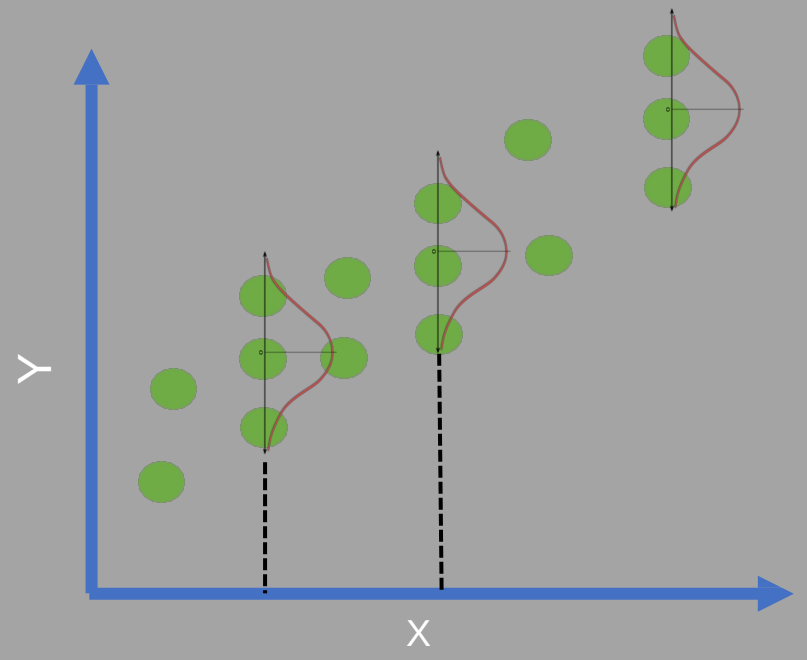


Homoscedasticity assumption

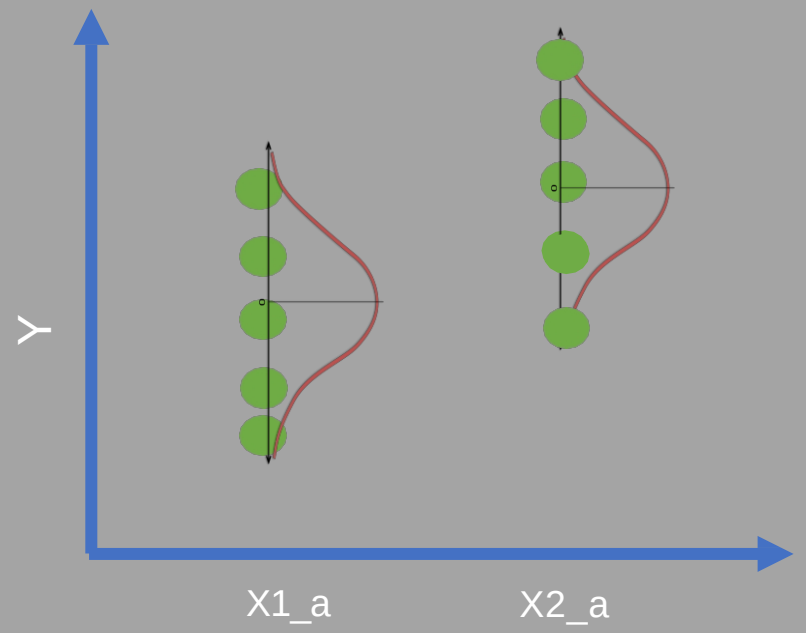


Normality assumption

Quantitative explanatory variable



Qualitative explanatory variable



Normality assumption

Too complicated/long, we go by residues.

What does residue normality mean?

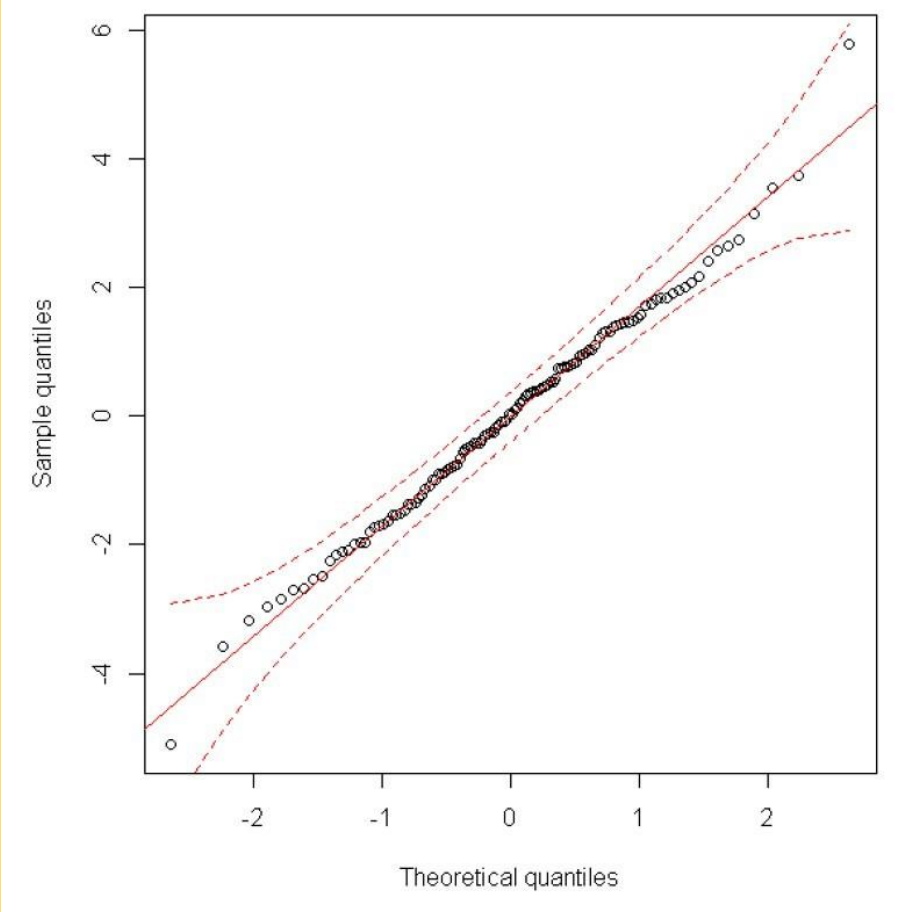
On average, I make small mistakes and as many large underestimates as large overestimates.

Normality assumption

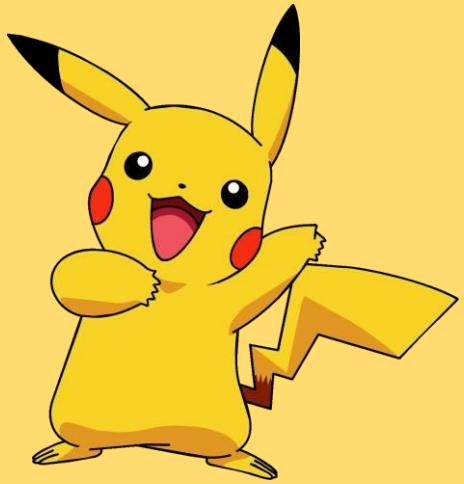
- Shapiro test on residuals (should not be significant)

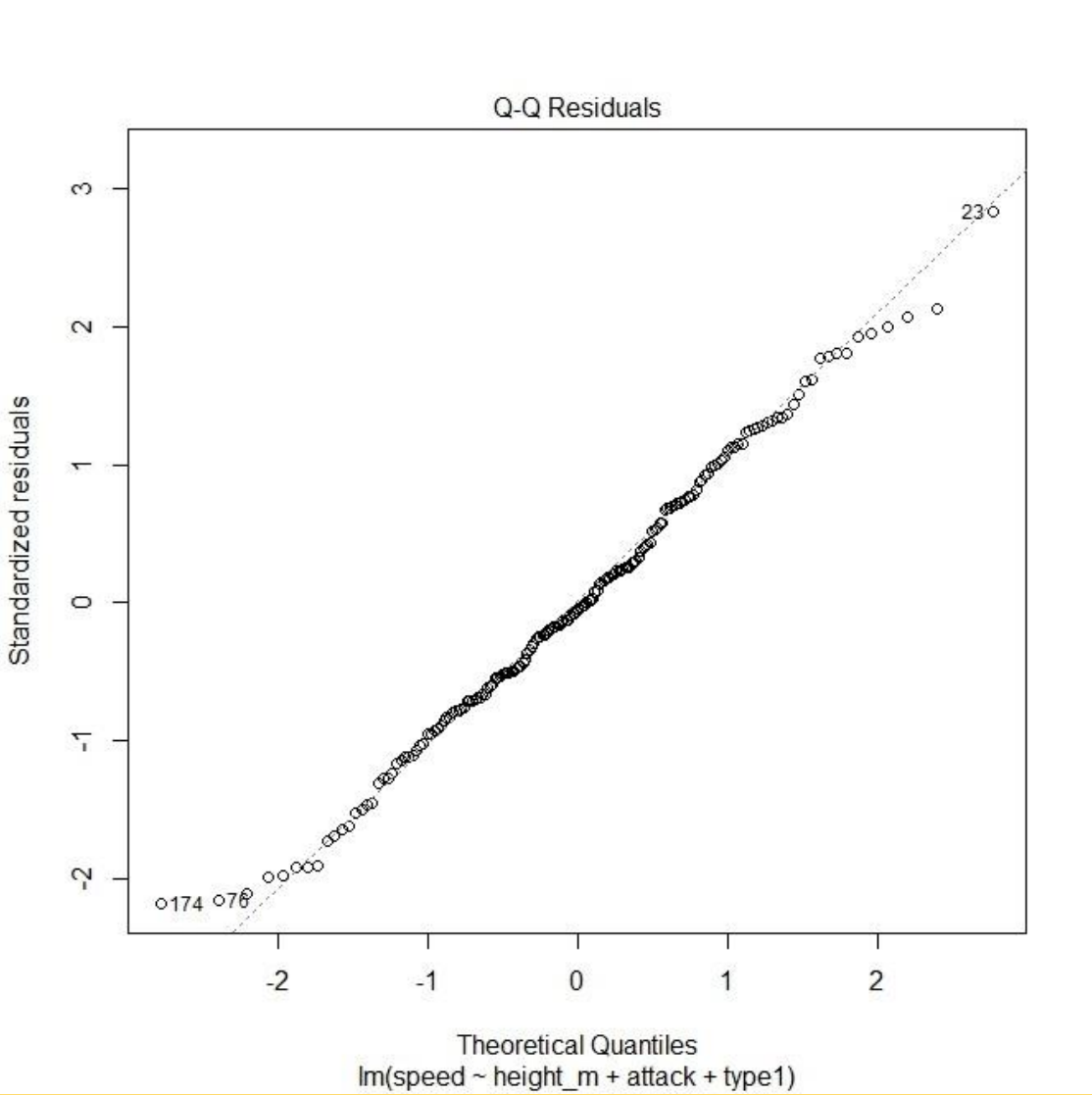
- qqplot

(the line must be globally on the 1st bisector and ~80% of the points must be aligned)

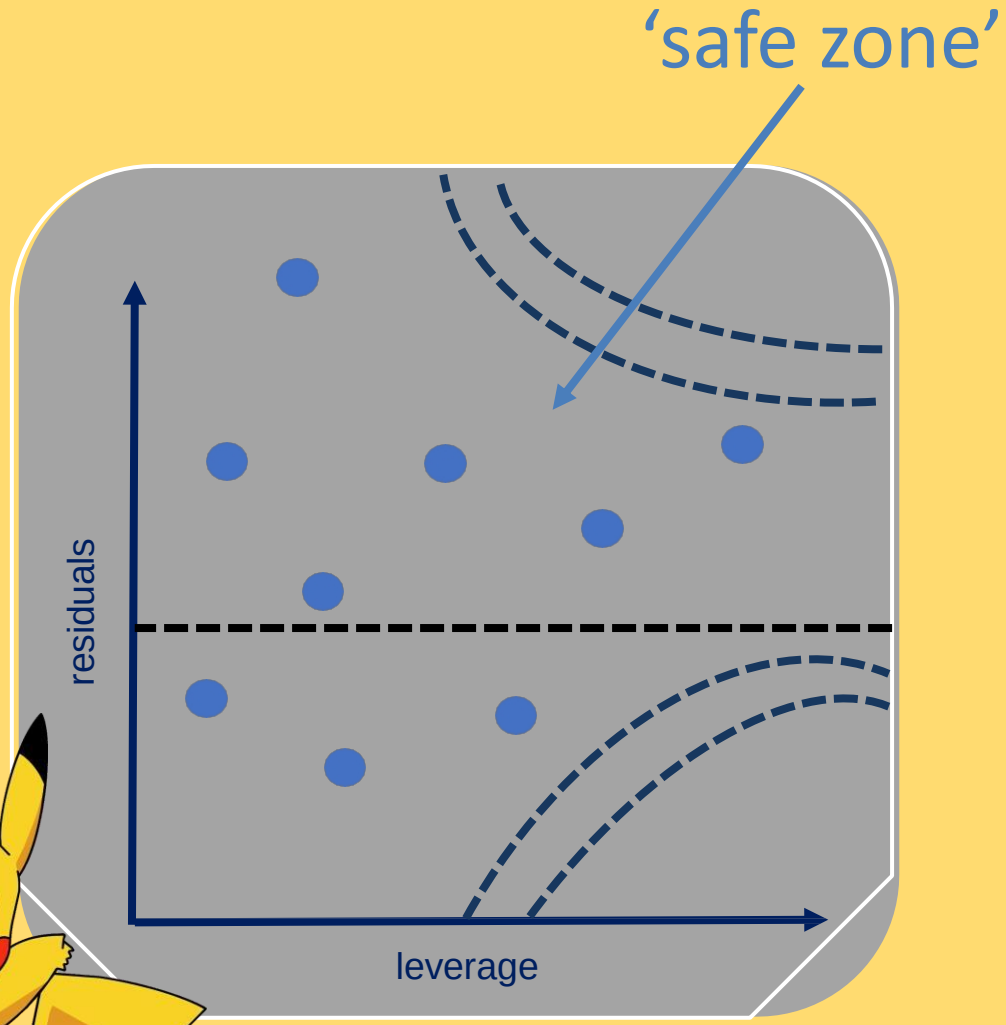


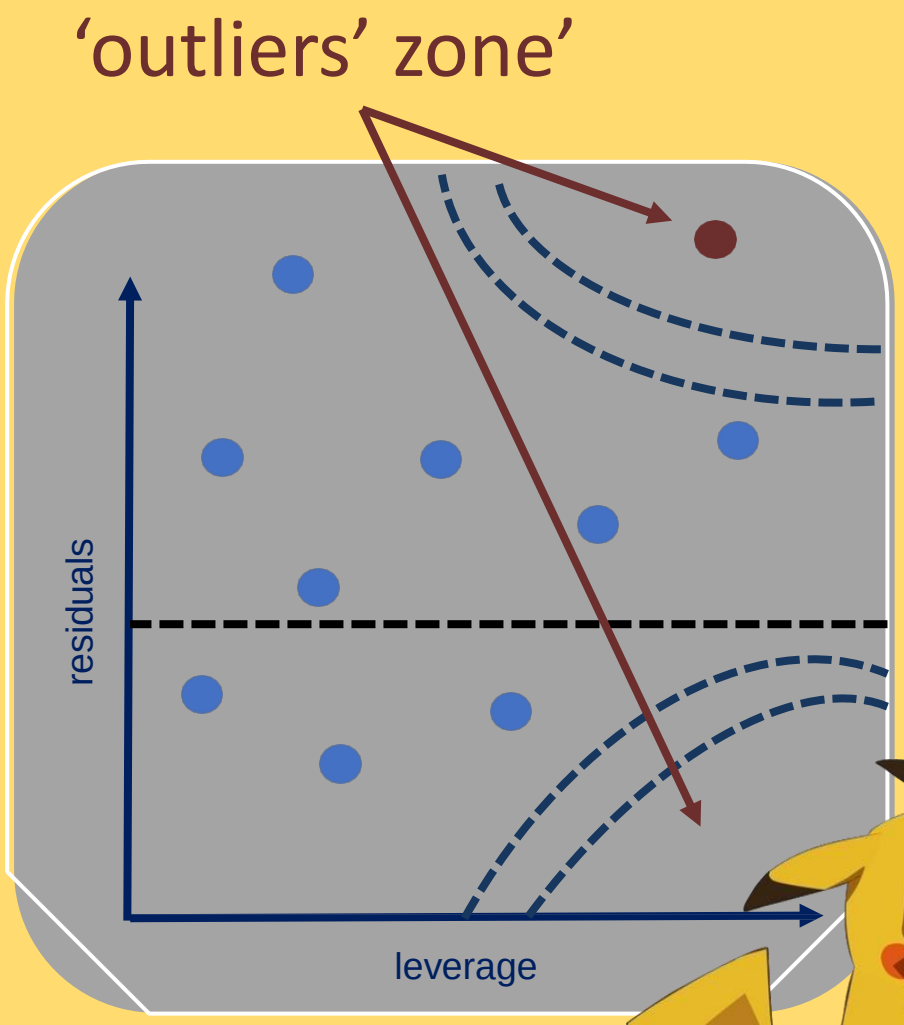
Normality assumption

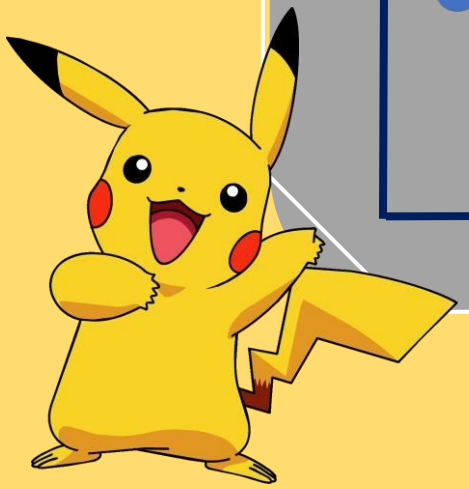




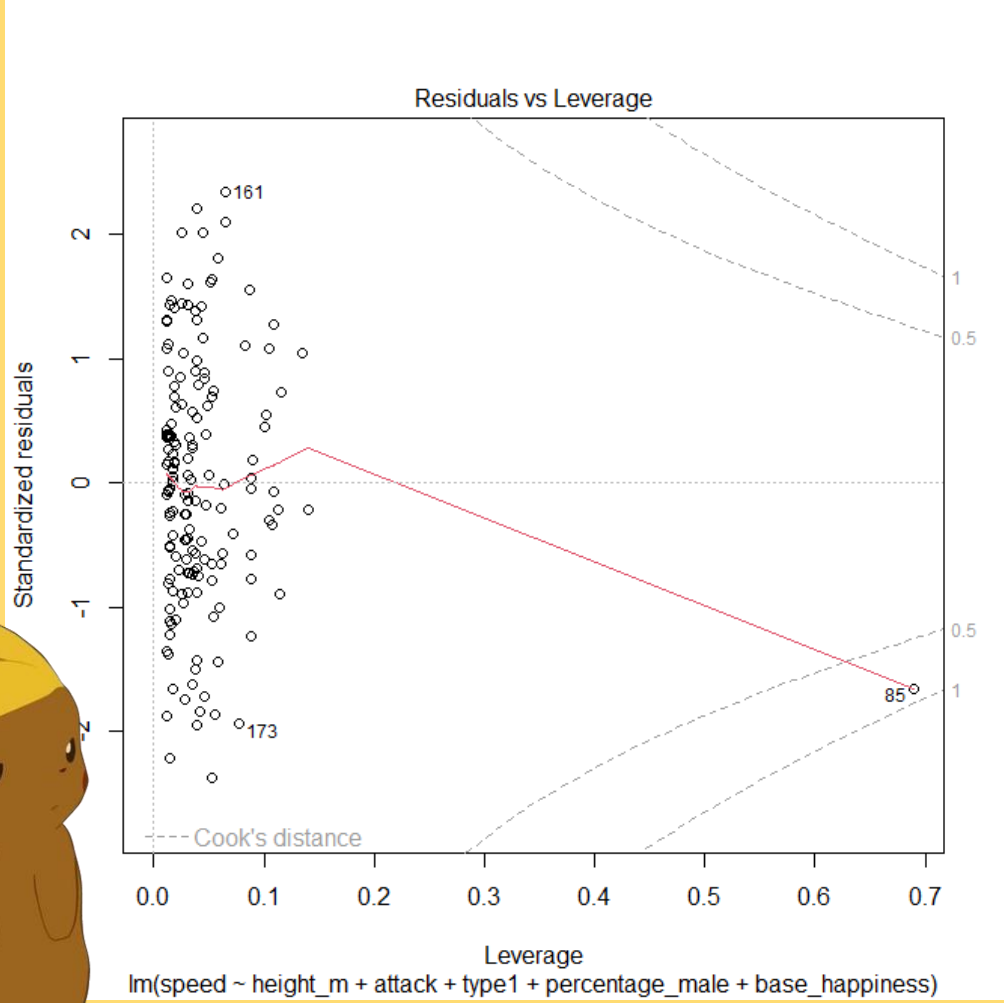
Outliers

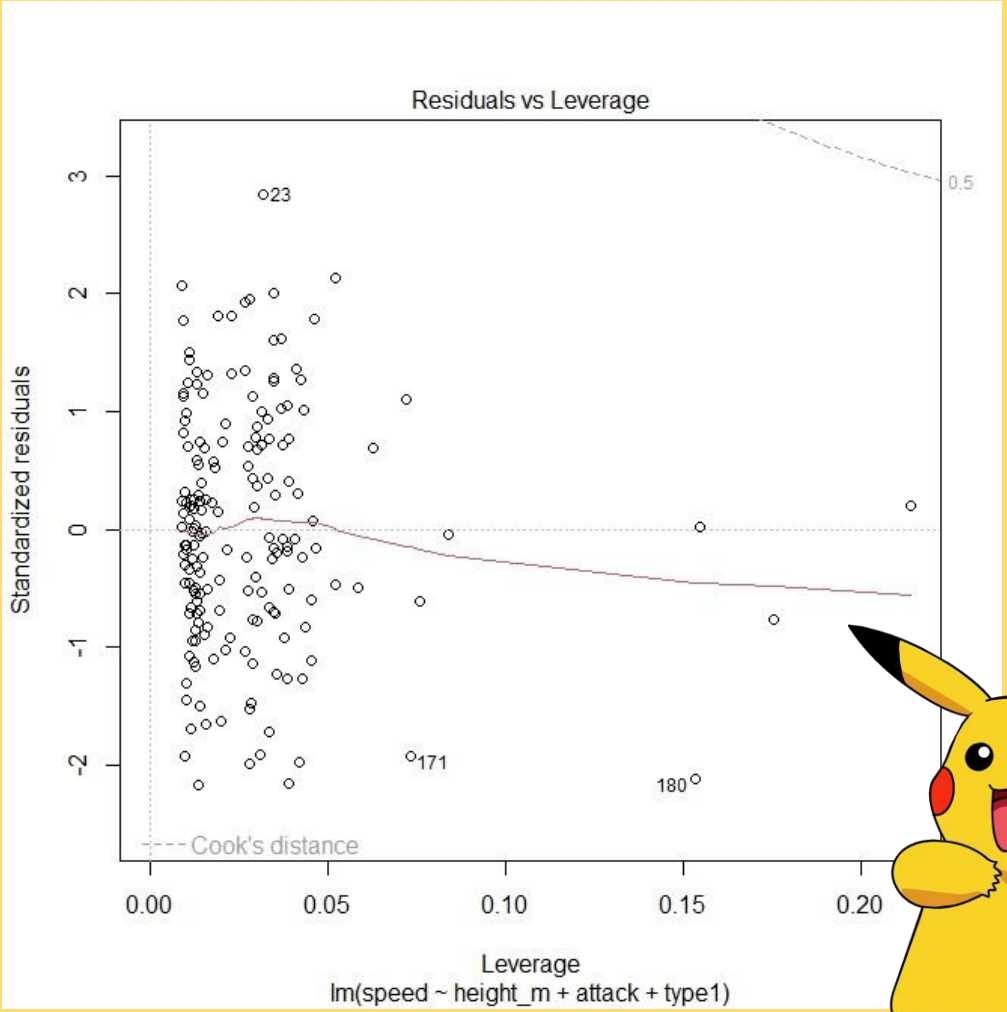

'safe zone'
residuals
leverage


'outliers' zone'
residuals
leverage

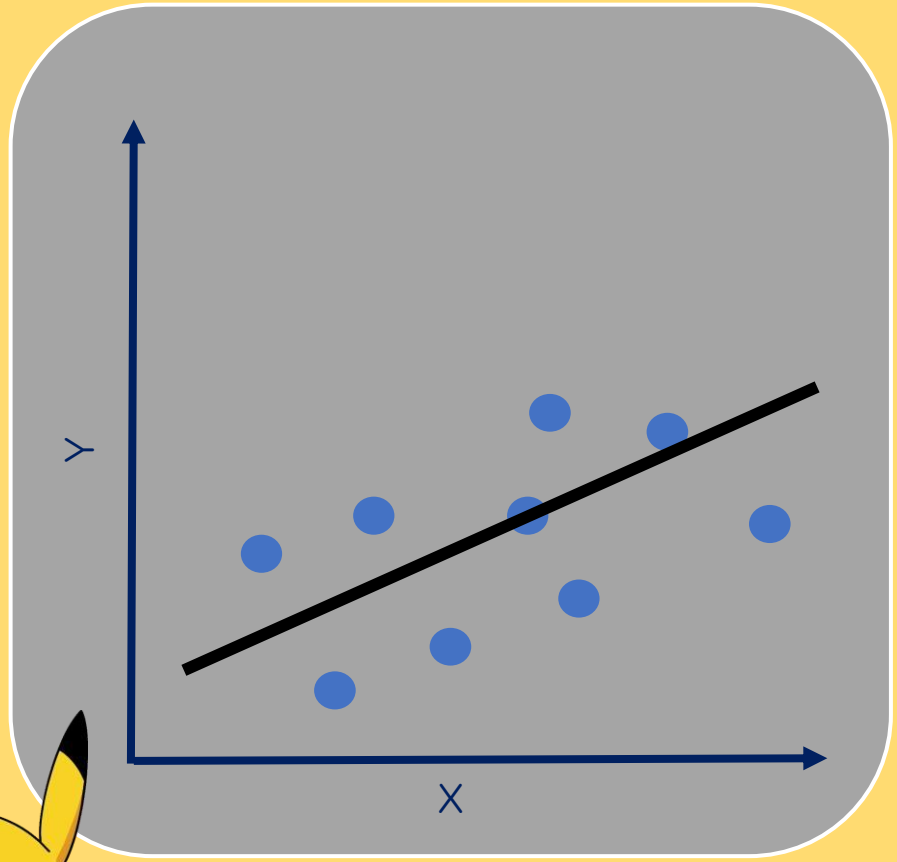


Outliers in Pokemon dataset

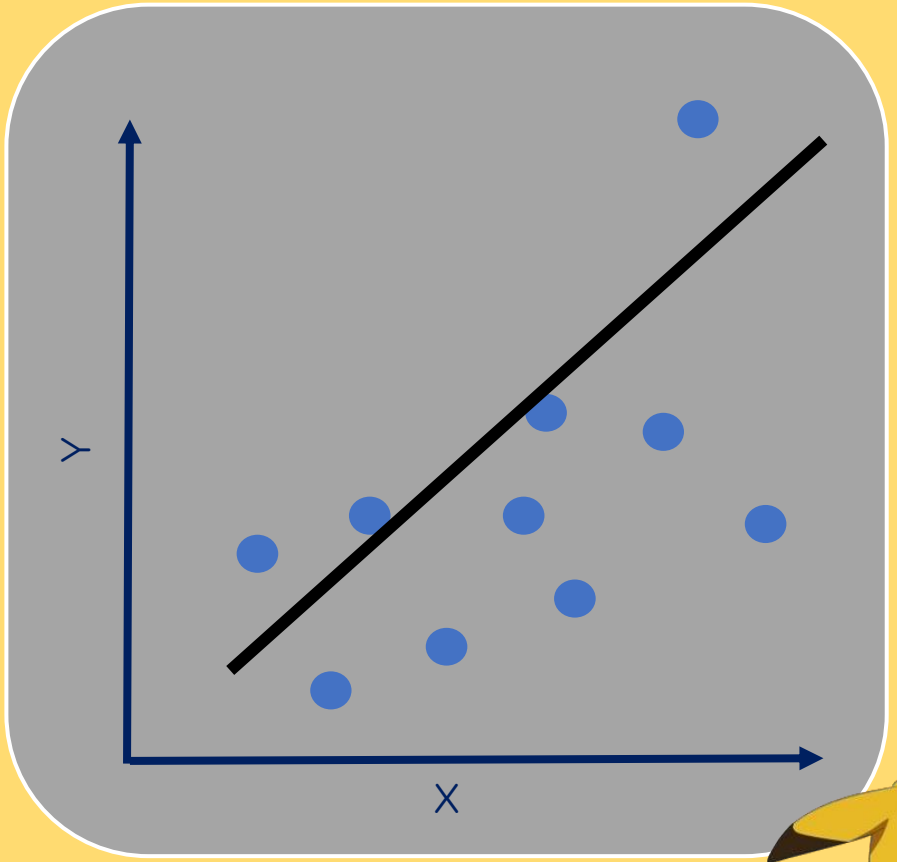




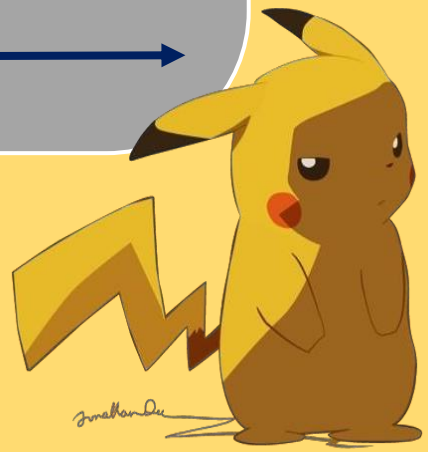
Outliers : leverage



No outliers
No leverage effect



One outlier
Leverage effect



The other checks are made on "residuals!

So not before the model !

Wrong reviewer comment :

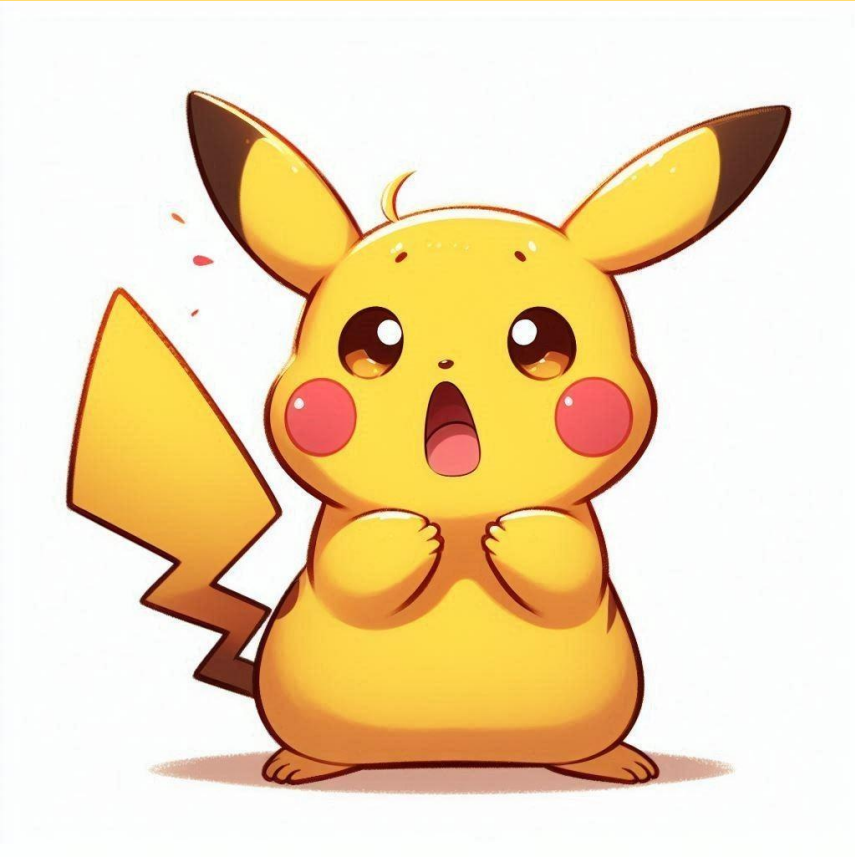
“Data distribution testing should perhaps already precede the linear models.”

Is it dramatic if the conditions for applying the models are not respected?

- Depends on domain

- It's hard to give a single answer: it depends on the hypothesis in question, the amplitude of the deviations and the number of data points.

- In general :
  - A slight linearity error is no big deal
  - The non-independence of residues is highly problematic
  - A lack of normality is less serious than a lack of homogeneity



Options?

- Non parametric tests :

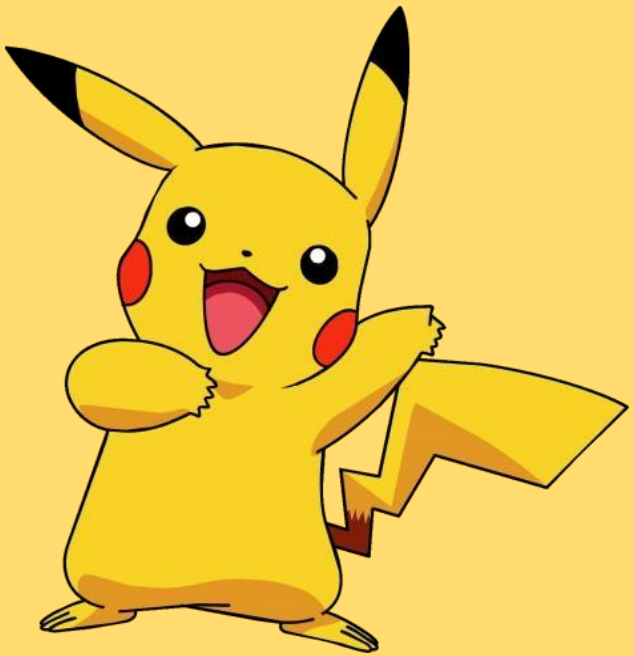
- Quanti-quanti: sperman
- Quanti-quali (2 modalities): wilcoxon
- Quanti-quali (>2 modalities): kruskall

- Permutation tests

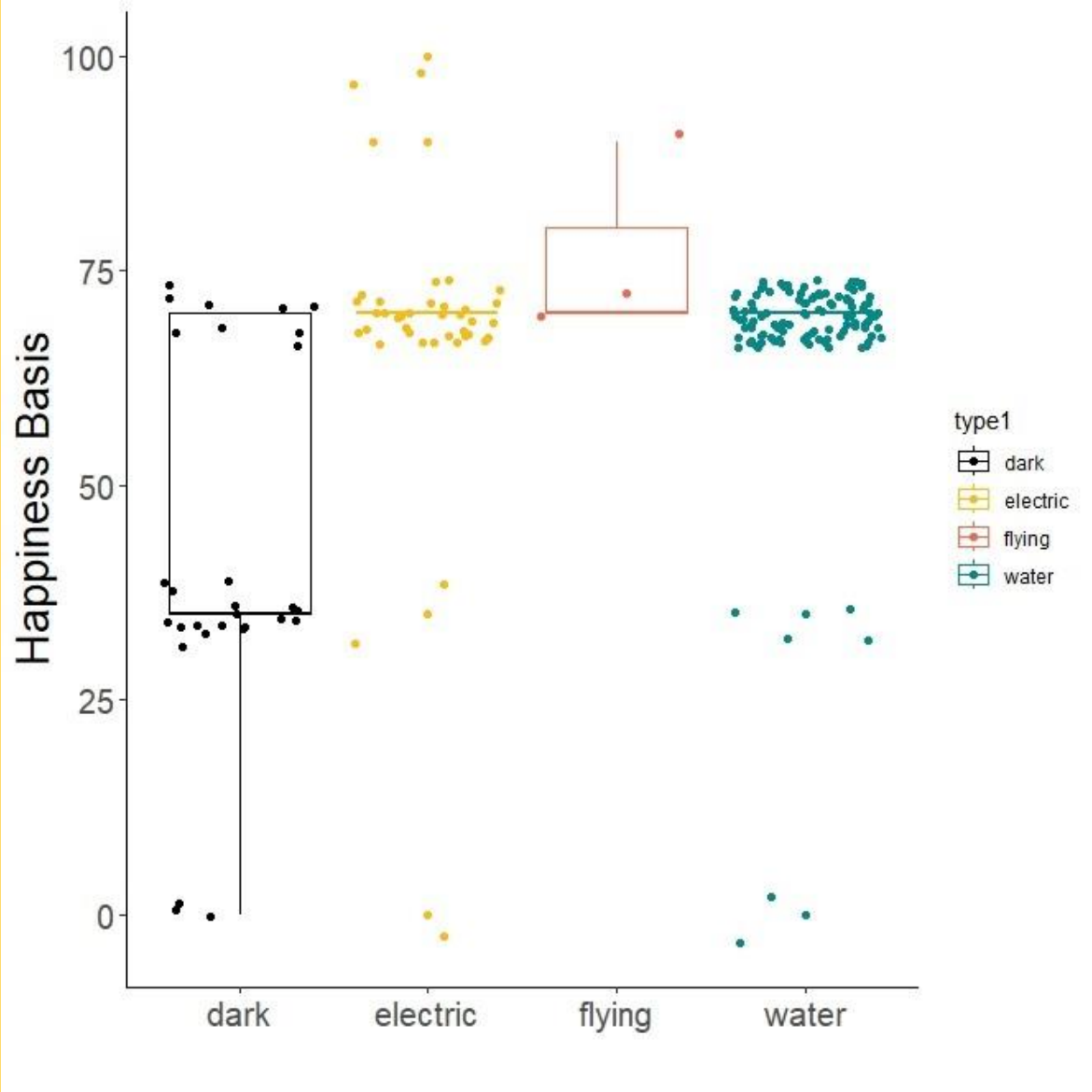
- Generalized Linear Models (GLM)

What happens with poor quality data and poor quality model?

What make a Pokemon Happy?

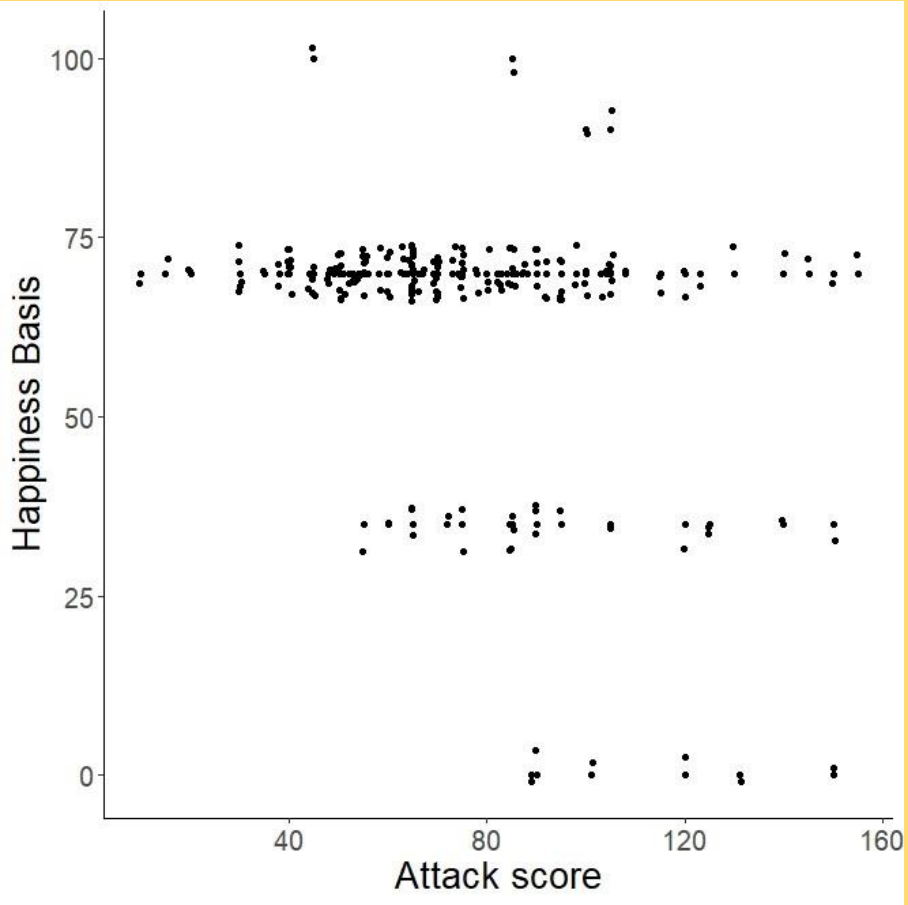


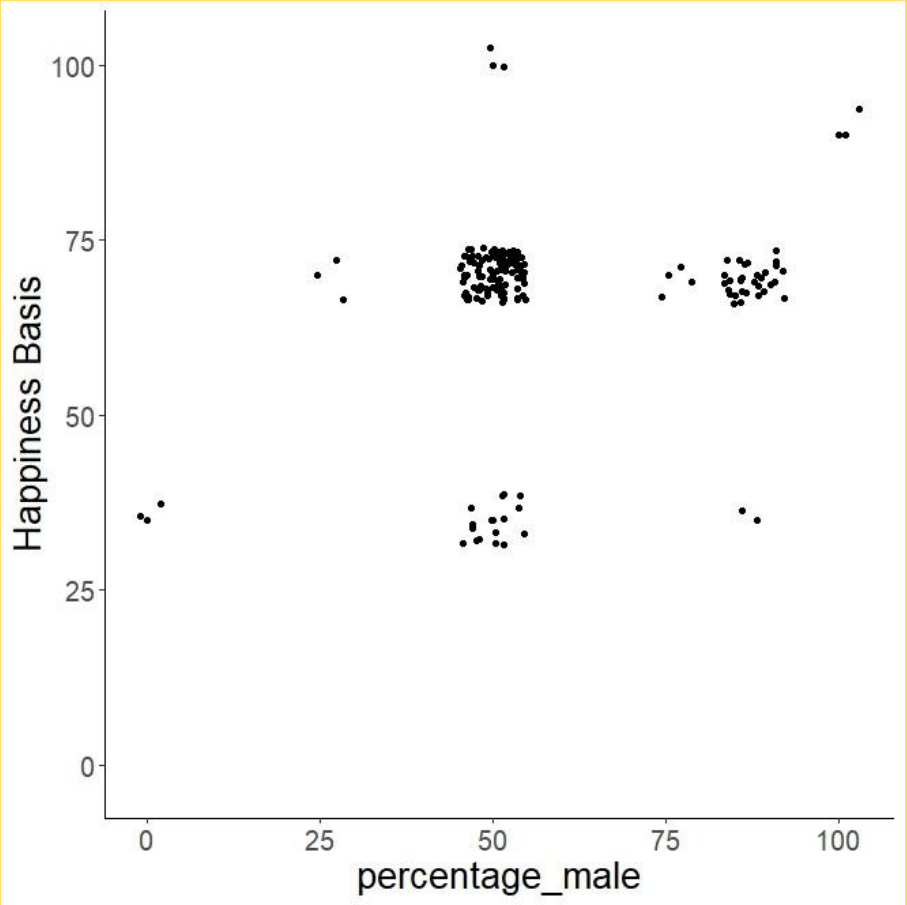
Data exploration



What do you think ?

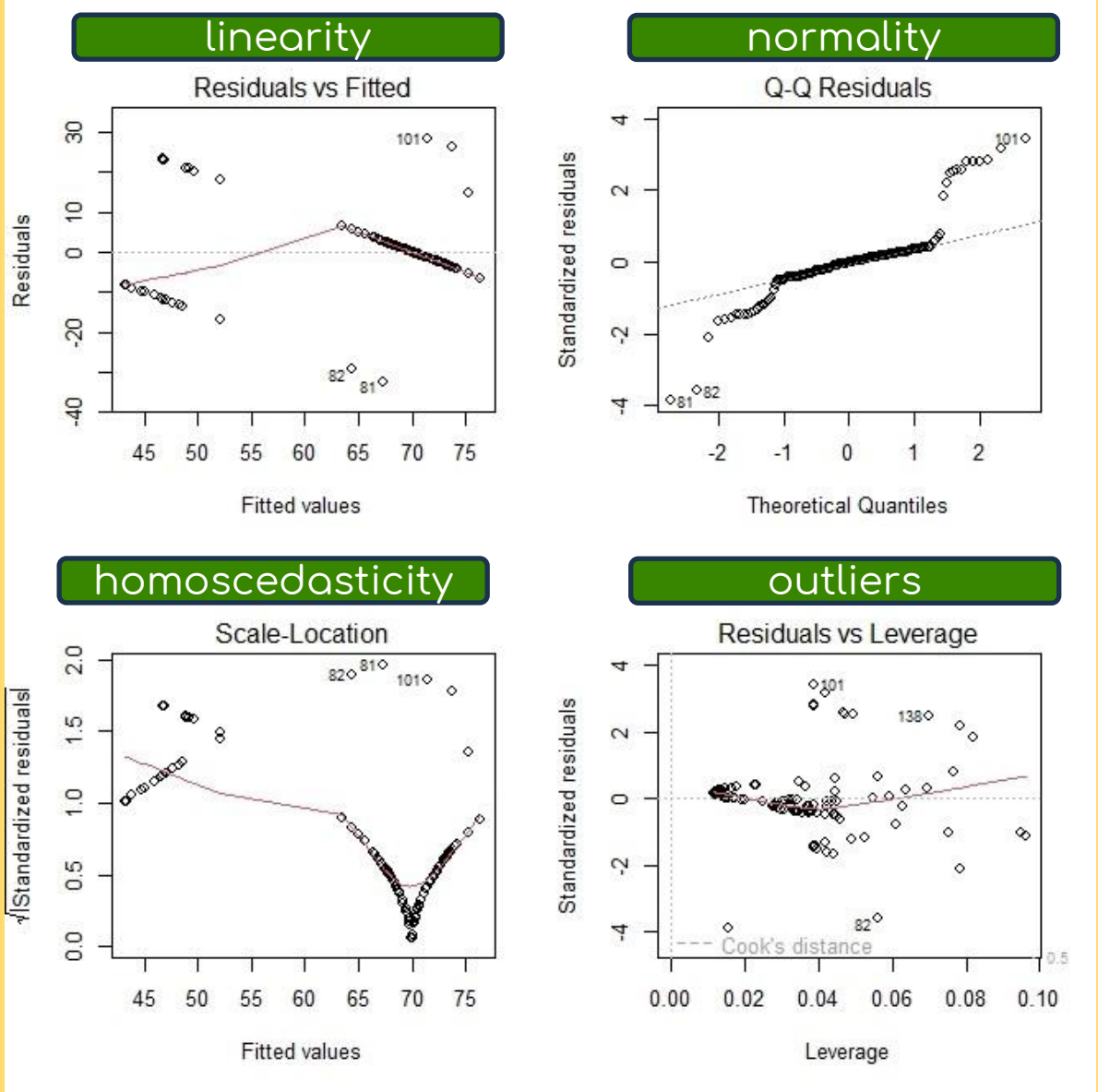
Data exploration





Cloud of data points ?

Application conditions



What do you think ?

Interpretation

```
> Anova(mod2)
Anova Table (Type II tests)

Response: base_happiness
                 Sum Sq  Df F value  Pr(>F)
percentage_male   431.7   1  6.0643 0.01490 *
attack            392.5   1  5.5135 0.02015 *
type1            9510.7   2 66.7936 < 2e-16 ***
Residuals       10892.8 153
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation

```
> summary(mod2)

Call:
lm(formula = base_happiness ~ percentage_male + attack + type1,
    data = data4)

Residuals:
    Min      1Q  Median      3Q     Max
-32.329  -2.865  -0.042   1.790  28.548

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     47.12270    3.12575  15.076   <2e-16 ***
percentage_male  0.09925    0.04030   2.463   0.0149 *
attack          -0.05874    0.02501  -2.348   0.0202 *
type1electric   24.36005    2.34296  10.397   <2e-16 ***
type1water      20.53039    1.92728  10.653   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.438 on 153 degrees of freedom
  (22 observations effacées parce que manquantes)
Multiple R-squared:  0.5327,    Adjusted R-squared:  0.5205
F-statistic:  43.6 on 4 and 153 DF,  p-value: < 2.2e-16
```

Interpretation

```
> tcm=emmeans(mod2,pairwise~type1)
> tcm
$emmeans
 type1    emmean    SE  df lower.CL upper.CL
 dark       48.5 1.711 153     45.1     51.8
 electric   72.8 1.597 153     69.7     76.0
 water      69.0 0.834 153     67.3     70.6

Confidence level used: 0.95

$contrasts
 contrast         estimate   SE  df t.ratio p.value
 dark - electric    -24.36 2.34 153 -10.397  <.0001
 dark - water       -20.53 1.93 153 -10.653  <.0001
 electric - water     3.83 1.80 153   2.125  0.0881

P value adjustment: tukey method for comparing a family of 3 estimates
```

The key stages of the linear model

Definition of the question

Exploration

Model construction

Checking application conditions

Interpretation and conclusion

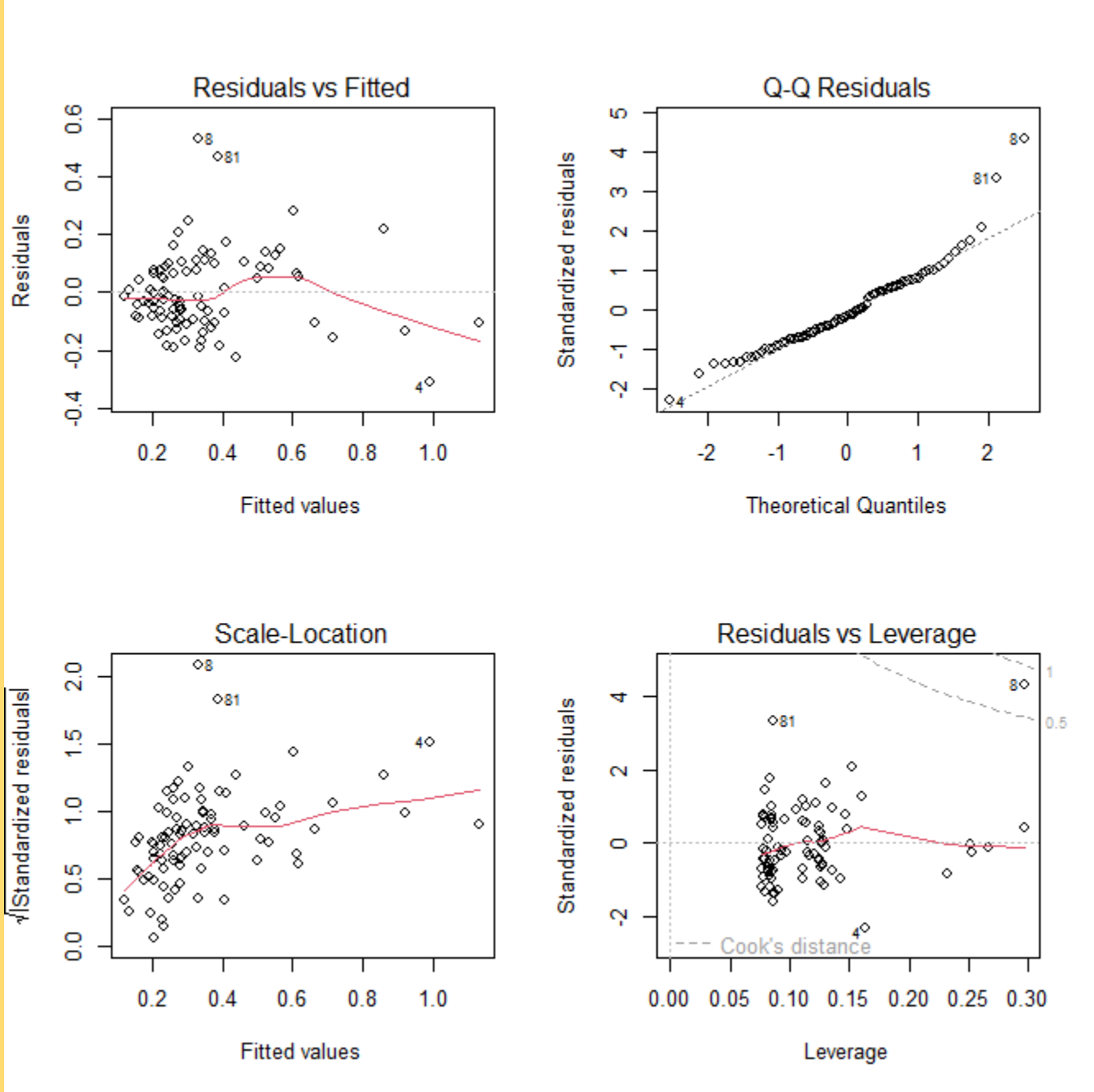
In your field

```
> mod=lm(Value1~Treatment+Value2,data)
```

```
> Anova(mod)
Anova Table (Type II tests)

Response: Value1
           Sum Sq Df F value    Pr(>F)
Treatment 0.29764  8   1.741    0.1023
Value2    1.16972  1  54.736 1.452e-10 ***
Residuals 1.64550 77
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In your field

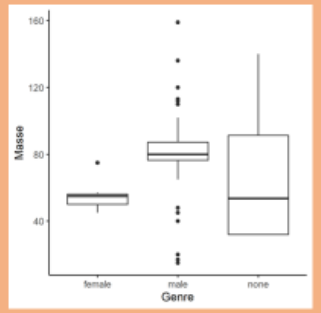
```
> summary(mod)

Call:
lm(formula = Value1 ~ Treatment + Value2, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.30654 -0.09529 -0.02493  0.08033  0.53210 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.0659499  0.0635353  -1.038   0.3025    
TreatmentD10 -0.0992562  0.0919577  -1.079   0.2838    
TreatmentD2  -0.0182919  0.0716861  -0.255   0.7993    
TreatmentD3   0.1039952  0.0675446   1.540   0.1277    
TreatmentD4   0.0627324  0.0658643   0.952   0.3438    
TreatmentD5   0.0437548  0.0730970   0.599   0.5512    
TreatmentD6   0.0681750  0.0661574   1.030   0.3060    
TreatmentD7   0.1137392  0.0667483   1.704   0.0924 .  
TreatmentNT  -0.1546060  0.1224403  -1.263   0.2105    
Value2        0.0026062  0.0003523   7.398 1.45e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1462 on 77 degrees of freedom
Multiple R-squared:  0.648,	Adjusted R-squared:  0.6069 
F-statistic: 15.75 on 9 and 77 DF,  p-value: 2.967e-14
```


Residuals vs Fitted
Residuals
Fitted values
Q-Q Residuals
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

Analysis

ANOVA analysis steps

Female | Male | None



Step 1: I build and run the model (here ANOVA)

```
mod<-lm(mass ~ sex , data=starwars)
```

Step 2: I recover the ANOVA table

```
anova(mod)
```

We recover a table that allows us to deduce whether the effects of each explanatory variable are significant or not. Absence of detail (i.e. which group has different averages from another).

Here, gender has a significant effect on mass (p = 0.041).

```
Analysis of Variance Table

Response: mass
          Df Sum Sq Mean Sq F value  Pr(>F)
sex        2   5346 2672.76  3.3835 0.04127 *
Residuals 54  42656  789.93
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Step 3: I retrieve the model summary table

```
summary(mod)
```

We recover the R^2, i.e. the % of variance (~ info) explained by the model.

We obtain certain "details". For the anova, we have the results of some 2-to-2 comparisons.

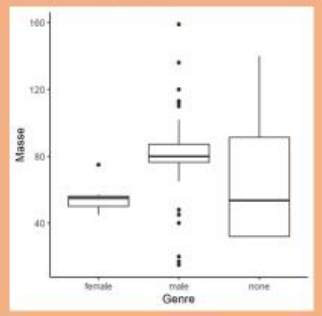
Here we have a significant difference in mass between the males and the reference modality (i.e. the one we don't see: female) (p=0.013). and no significant difference between none and the reference modality (p=0.377).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   54.689      9.369   5.837  3.1e-07 ***
sexmale       26.316     10.282   2.559   0.0133 *
sexnone       15.061     16.889   0.892   0.3765
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.11 on 54 degrees of freedom
  (25 observations effacées parce que manquantes)
Multiple R-squared:  0.1114,    Adjusted R-squared:  0.07845
F-statistic: 3.384 on 2 and 54 DF,  p-value: 0.04127
```

Step 4: I run multiple comparison tests if significant effects

Writing !

ANOVA writing




```
mod<-lm(mass ~ sex , data=starwars)
```

Anova table

```
anova(mod)
```

According to the anova table, gender has a significant effect on mass (ρ = 0.041).

```
Analysis of Variance Table

Response: mass
          Df Sum Sq Mean Sq F value  Pr(>F)
sex        2   5346 2672.76  3.3835 0.04127 *
Residuals 54  42656  789.93
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary output

```
summary(mod)
```

The p-value of the Fisher test is less than 0.05 (ρ=0.041), which means that the model is well explanatory.

The R^2 is 0.11, meaning that 11% of the variation in mass can be explained by gender.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   54.689      9.369   5.837  3.1e-07 ***
sexmale       26.316     10.282   2.559   0.0133 *
sexnone       15.061     16.889   0.892   0.3765
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.11 on 54 degrees of freedom
  (25 observations effacées parce que manquantes)
Multiple R-squared:  0.1114,    Adjusted R-squared:  0.07845
F-statistic: 3.384 on 2 and 54 DF,  p-value: 0.04127
```

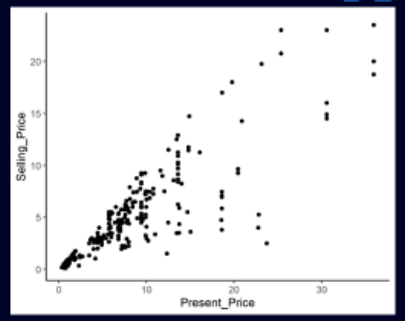
From the summary output, I can say that :

Males are significantly heavier than females (ρ=0.013).
Asexuals are significantly heavier than females (ρ=0.377).

Analysis

Regression interpretation

Does the present price influence the selling price of a car?



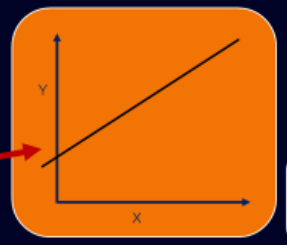
Step 1: I create and run the model(here simple regression)

```
mod=lm(Selling_Price~Present_Price,data)
```

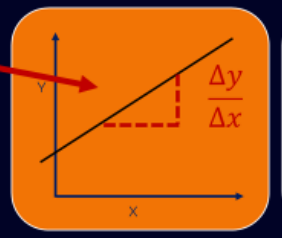
Step 2: I retrieve the model's summary table

```
summary(mod)
```

```
Call:
lm(formula = Selling_Price ~ Present_Price, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.3598  -0.4632  -0.1490   0.7554   8.1943 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    0.33841    0.18041   1.876   0.0617 .  
Present_Price  0.56980    0.01794  31.765   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.141 on 295 degrees of freedom
Multiple R-squared:  0.7738,	Adjusted R-squared:  0.773 
F-statistic:  1009 on 1 and 295 DF,  p-value: < 2.2e-16
```



Y-intercept :
Y value when X is 0.



$$\frac{\Delta y}{\Delta x} = 0.570$$

X increases of 1 unit, $\Delta x = 1$ so $\Delta y = 0.570$. slope coefficient corresponds to what 'we win' on y for a 1-unit x increase.

Note: between steps 1 and 2, we could have retrieved the anova table

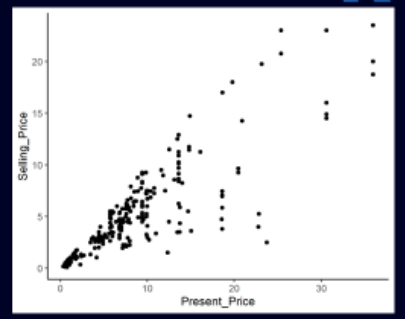
```
> anova(mod)
Analysis of Variance Table

Response: Selling_Price
               Df Sum Sq Mean Sq F value    Pr(>F)    
Present_Price   1 4626.8  4626.8    1009 < 2.2e-16 ***
Residuals     295 1352.7     4.6                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

It's not wrong to want to retrieve the anova table when doing a regression. It's just that it's not very informative. In the summary we already have one line per variable, which allows us to tell whether this variable has a significant effect.

Writing !

Rédaction régression linéaire

Does the present price influence the selling price of a car?



Step 1: interpretation of the p-value and R^2 of the model

```
summary(mod)
```

```
Call:
lm(formula = Selling_Price ~ Present_Price, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.3598  -0.4632  -0.1490   0.7554   8.1943 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    0.33841    0.18041   1.876   0.0617 .  
Present_Price  0.56980    0.01794  31.765   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.141 on 295 degrees of freedom
Multiple R-squared:  0.7738,	Adjusted R-squared:  0.773 
F-statistic:  1009 on 1 and 295 DF,  p-value: < 2.2e-16
```

The p-value of the Fisher test is less than 0.05, which means that the model is well explanatory.

The R^2 is 0.77, which means that 77% of the variation in sales price can be explained by the current price.

Step 2: Interpretation of model coefficients

Slope

The coefficient associated with current price is significantly different from 0 ($p < 0.001$).

There is a significant effect of current price on selling price.This coefficient is positive. When the current price increases, the selling price also increases.

If the current price is increased by one unit (e.g. 100 euros), the selling price increases by 0.57 units (e.g. 57 euros).

Y-intercept

The y-intercept is not significantly different from 0 ($p = 0.062$).

When the current price is 0, the selling price is also 0.