Recall: ggplot()- layered grammar of graphics

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(
     mapping = aes(<MAPPINGS>),
     stat = <STAT>,
     position = <POSITION>
  ) +
  <COORDINATE_FUNCTION> +
  <FACET_FUNCTION>
```

Key tool in Data Visualisation

Coordinate chart

Scatter plot

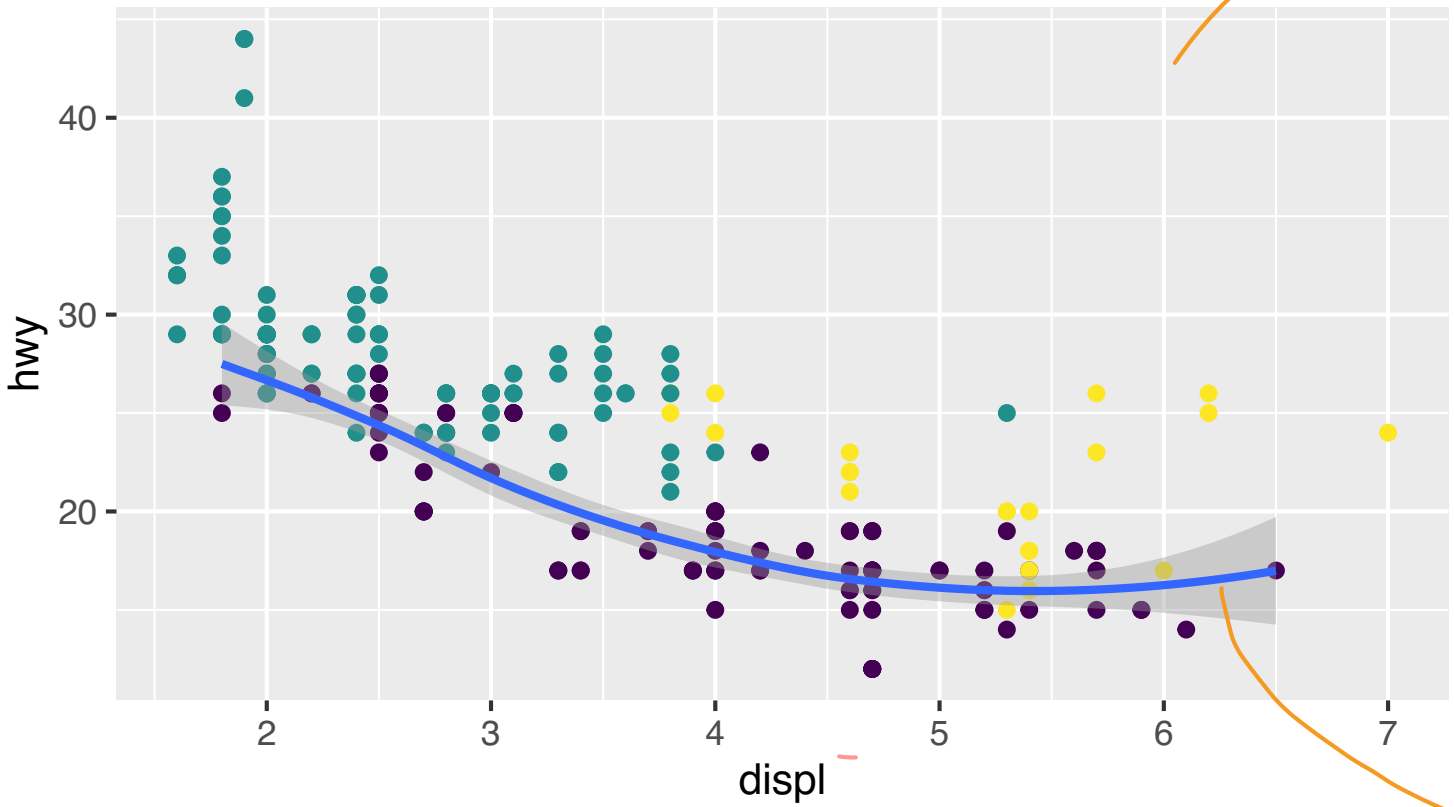
Smooth line

bar chart

polar coordinates

Faceting.

ggplot()-geoms

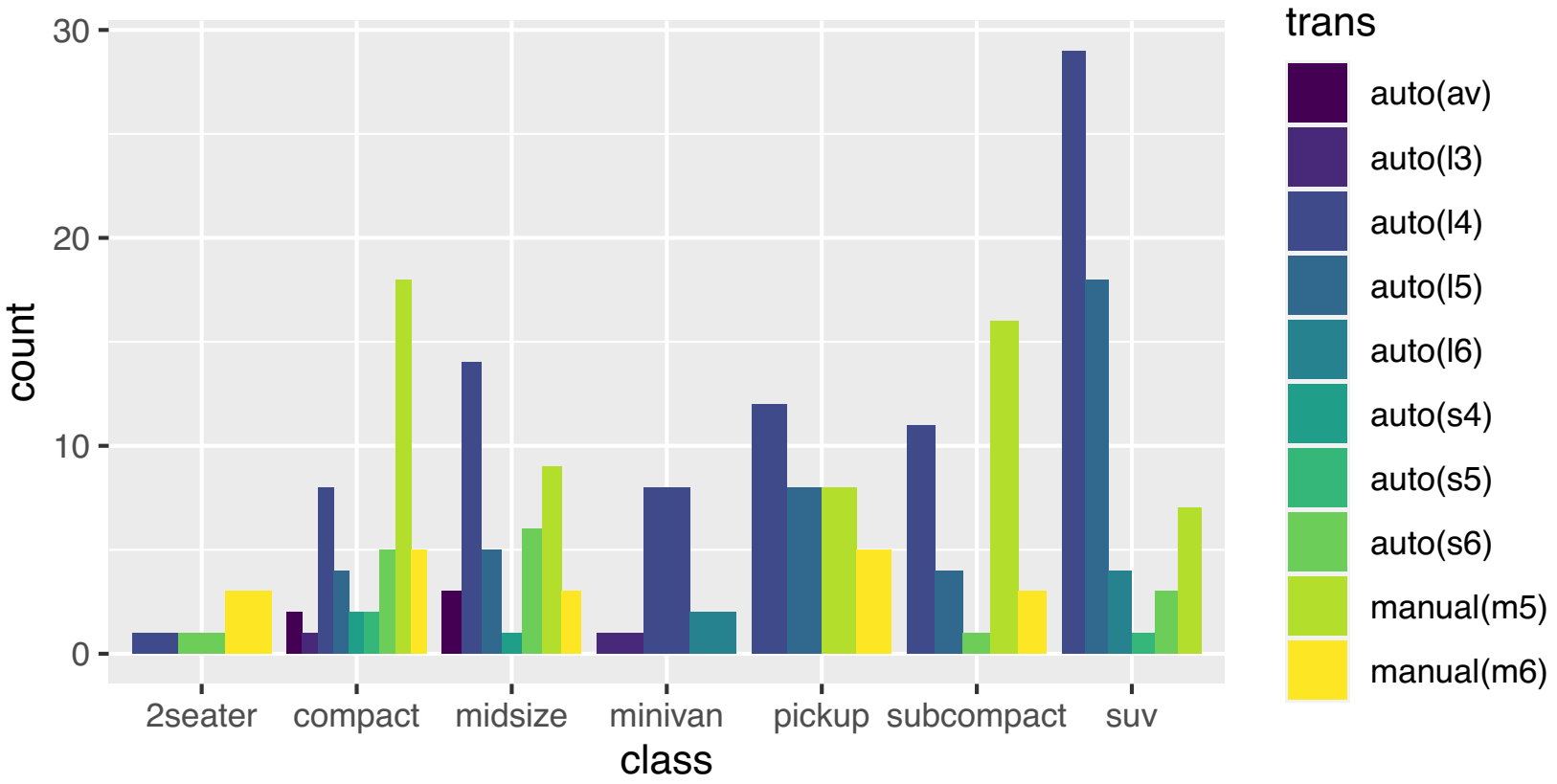


4, f, r - drive
colours

geon - smooth

filter: 4 wheel drive

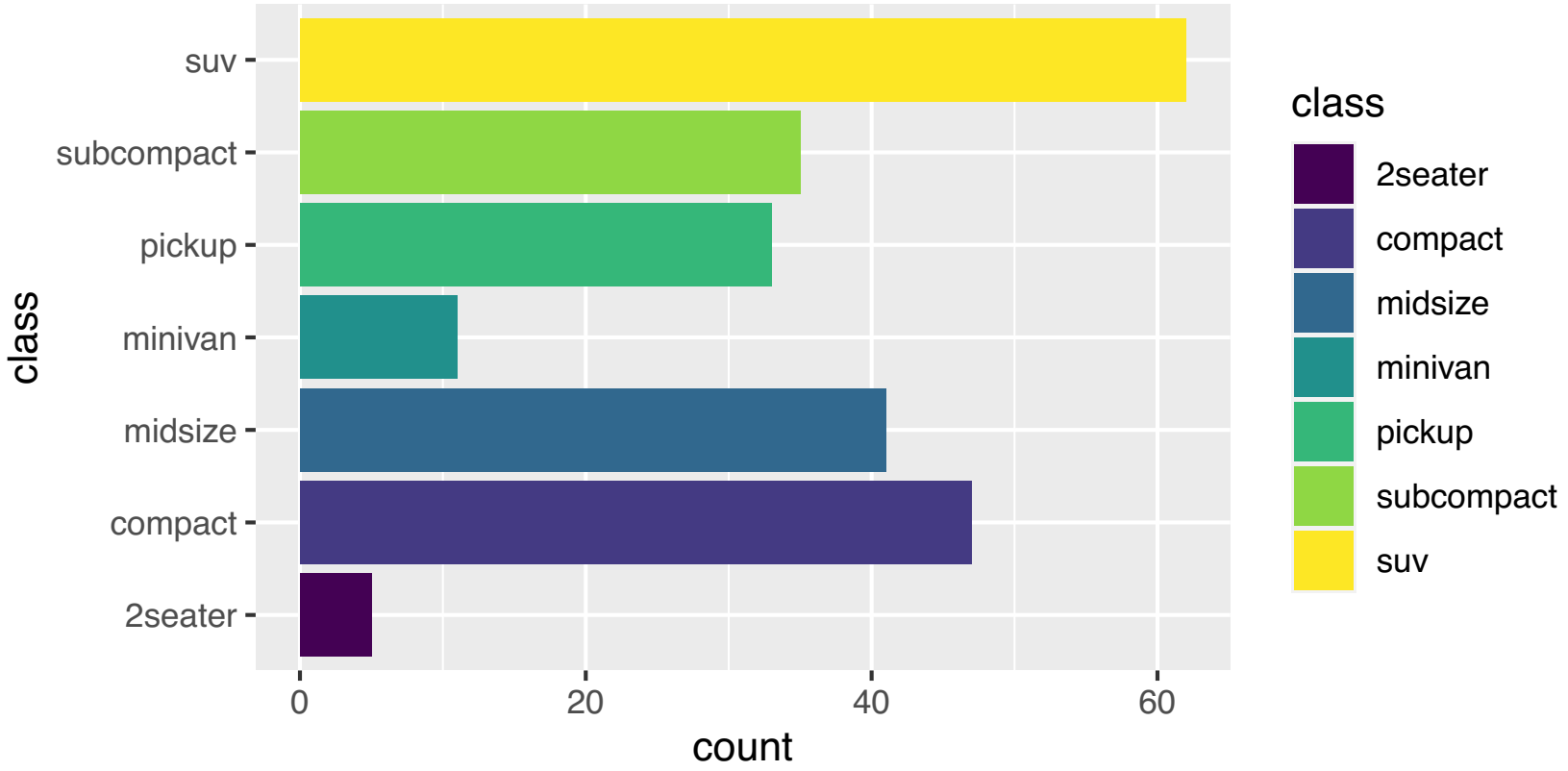
Dodge - Position



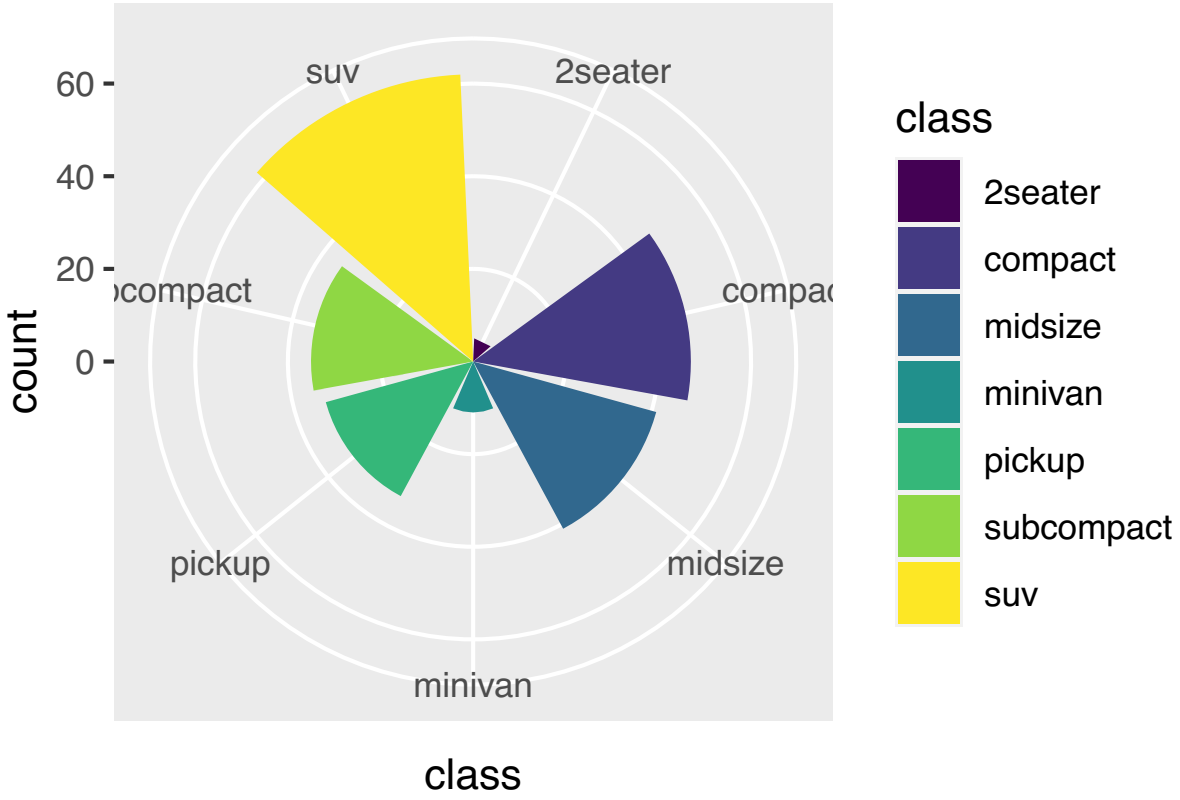
ggplot()-barchart

. Vinds fill ≡ class.

. Horizontal



Polar Coordinates.



Model: Aims and Methods

- No statistical Analysis in today's classes

- introduce
  - linear model
  - (Exploratory)

how it works ?

- Datasets & week 2 R Code —— Please download the data sets & R-code.

Website:https://www.isibang.ac.in/~athreya/Teaching/LSM/

- while working with the R-code
- Please set Correct working directory

(Place R-code and data set in the same directory)

Aim:- To find relationship between y and x.

Given: Data set

x - independent

y - dependent on x } variable

x	y
x_1	y_1
x_2	y_2
⋮	⋮
x_{100}	?

Goal (today):

- low-dimensional summary of the data set.
- "patterns" in data
- prediction ?

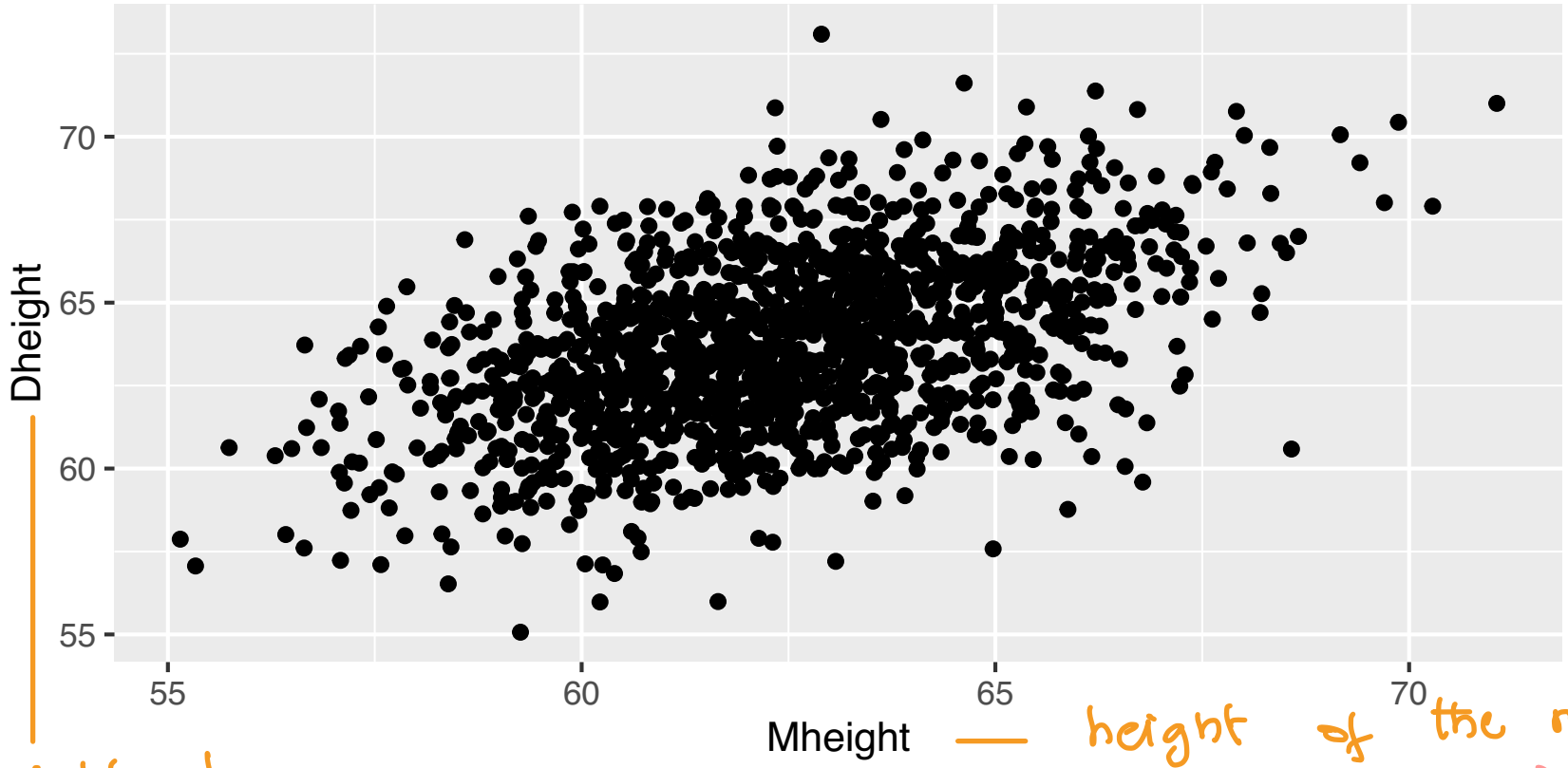
Caution:-

All models are wrong but some are useful.

- $PV = RT$ - model for ideal gases
  - good approximation to understand the system

1893-98 : K.S. Pearson collected ~ 1500 odd

(heights.txt)



jitter plot

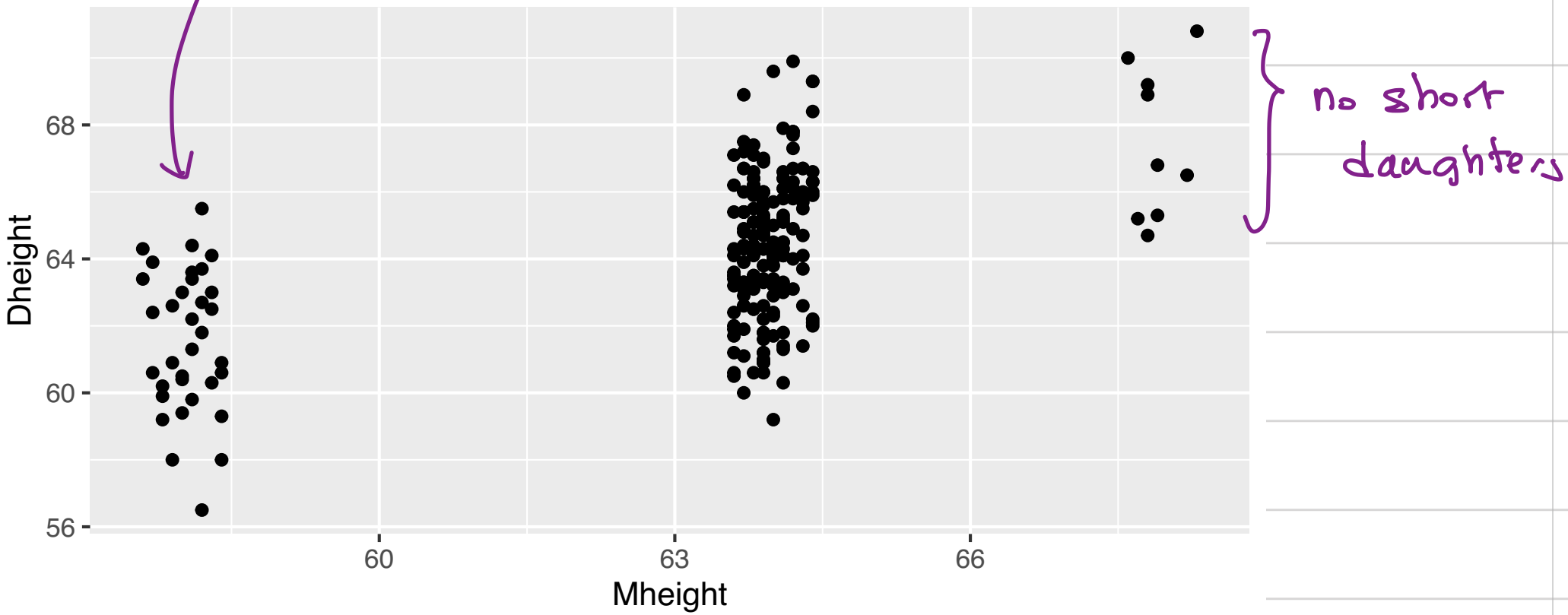
moves each data point by a small uniform r.v.

height of one adult daughter (age > 18)

height of the mother (age < 65)

Question :- How does the daughter inherit the height of the mother?

object of study



identify outliers.

- Develop a technique to use a linear model

$\{(x_i, y_i) : 1 \le i \le n\}$ - Data

Find: a_1, a_2 :

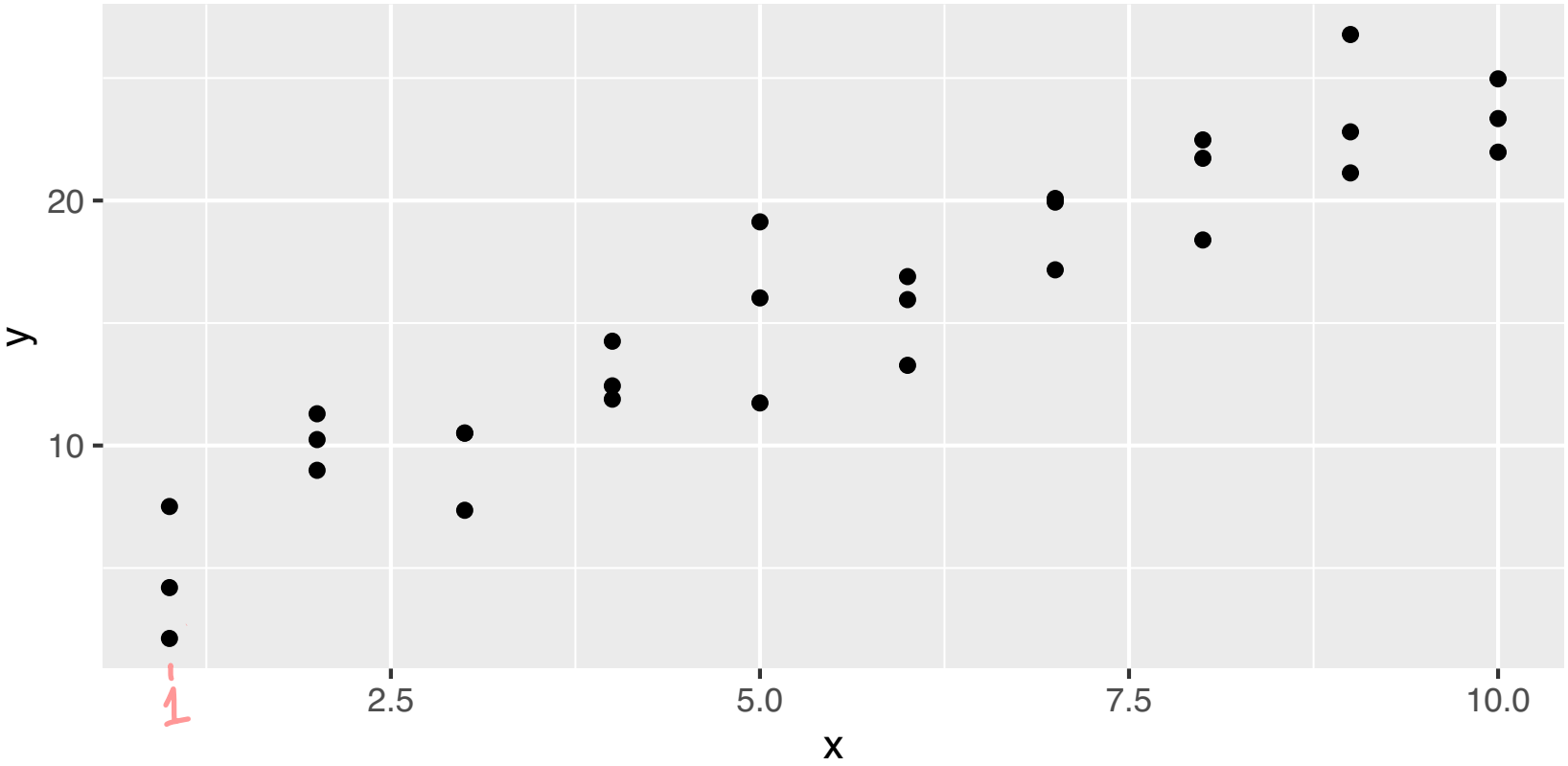
$y = a_2 x + a_1$ is a good approximation

In practice:

- what are hypotheses one wants to test for the data?
- Training — 60% — 1 or 2 models
- Query — 20% — which works better
- Test — 20% — Performance evaluation

- Work with Simulated data set
  - package called "modelr"
- sim1 - data set that we will use.

- sim1 – data set

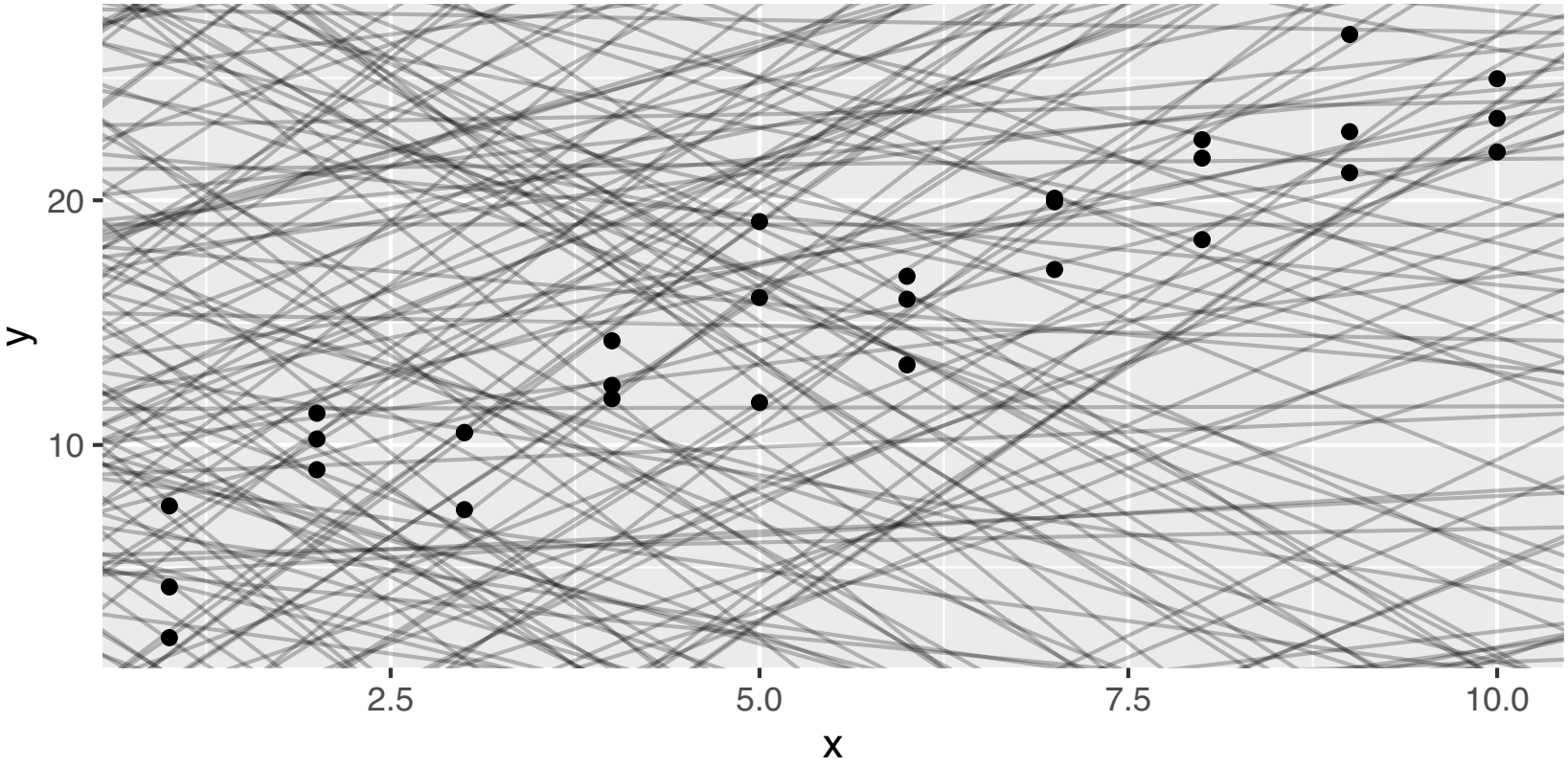


- There is a pattern

- "linear relationship"

$a_1 \equiv$ intercept – 250 uniformly chosen r.v. in $[-20, 40]$

$a_2 \equiv$ slope – 250 uniformly chosen r.v. in $[-5, 5]$



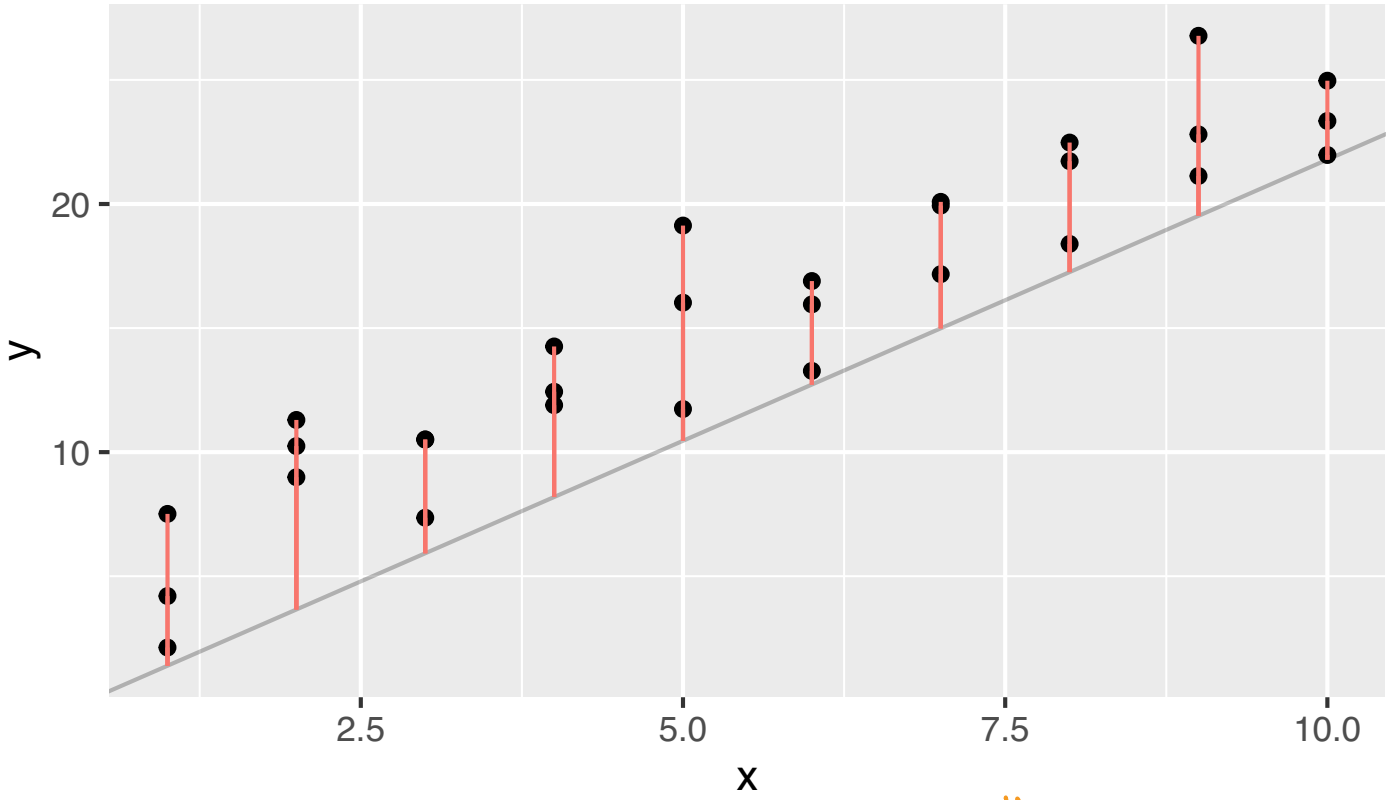
Plotted the 250 lines & data set

250 linear models

Linear Models : Understand Errors

- Chosen 130th line
- plotted the line

```
[1] 130
```



Residuals are the distance of the points from the line.

Criteria for good model :- minimize the "sum of distances" (X)
the "sum of |distances|" (Hard)
the "sum of (distances)^2" ✓ least square"

Linear Models : find the best possible line

Data: $\{(x_i, y_i) : 1 \le i \le n\}$

measure-distance (a_1, a_2)

$$= \left[\frac{1}{n} \sum_{i=1}^{n} [y_i - (a_1 + a_2 x_i)]^2 \right]^{\frac{1}{2}}$$

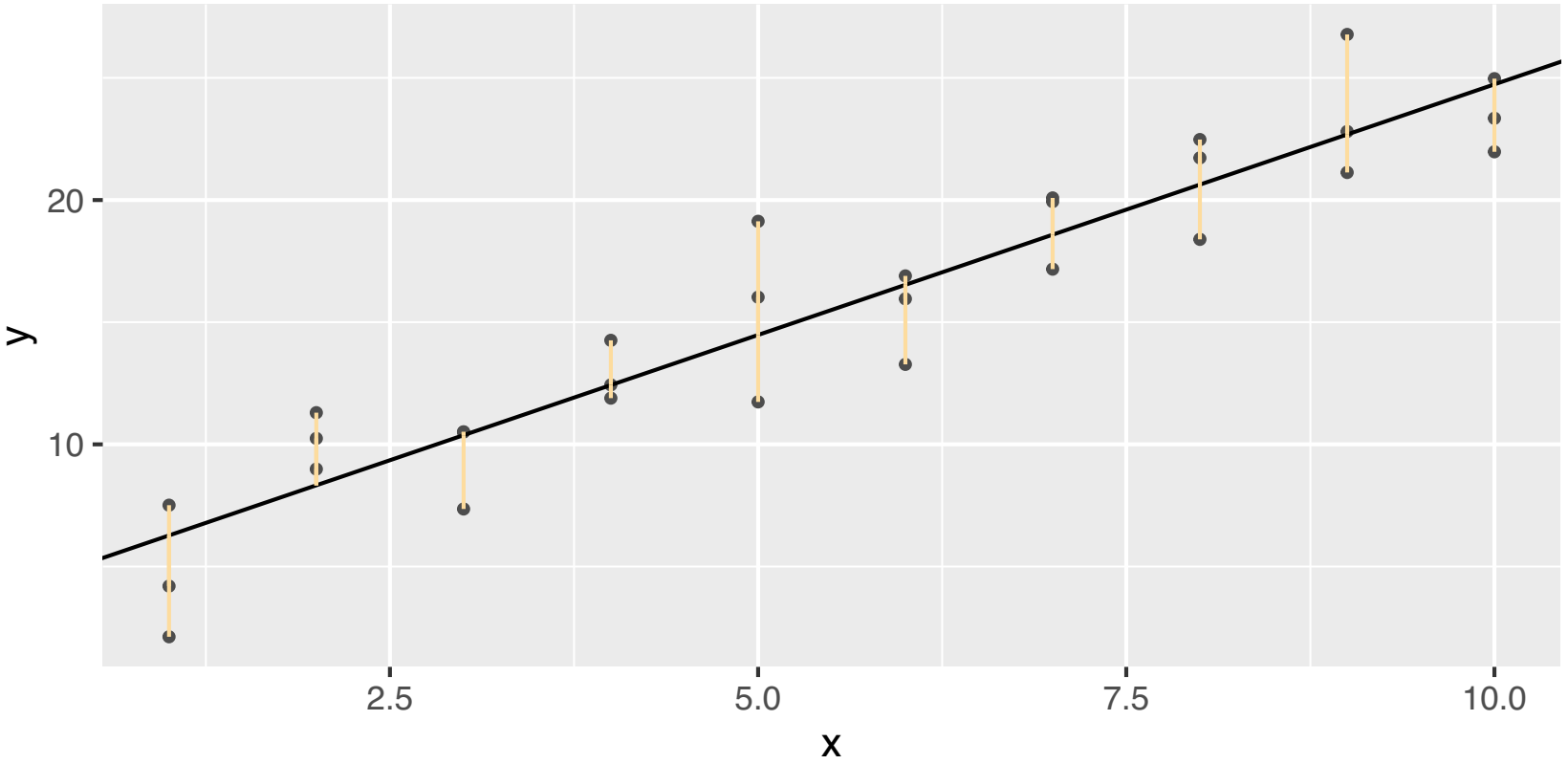
```
> model1 <- function(a, data) {
+   a[1] + data$x * a[2]
+ }
> #measure distance
> measure_distance <- function(mod, data) {
+   diff <- data$y - model1(mod, data)
+   sqrt(mean(diff ^ 2))
+ }
> best <- optim(c(0, 0), measure_distance, data = sim1)
> best$par
[1] 4.222248 2.051204
> #> [1] 4.222248 2.051204
> measure_distance(c(best$par[1], best$par[2]) , sim1)
[1] 2.128181
```

Find the (a_1, a_2) that minimizes:

measure-distance (a_1, a_2)

Linear Models : best line

Plotted the line : intercept :- 4.22, Slope :- 2.05



Seems to provide a good approximation

- How to judge the model ?
- Is this the best possible solution under the model ?

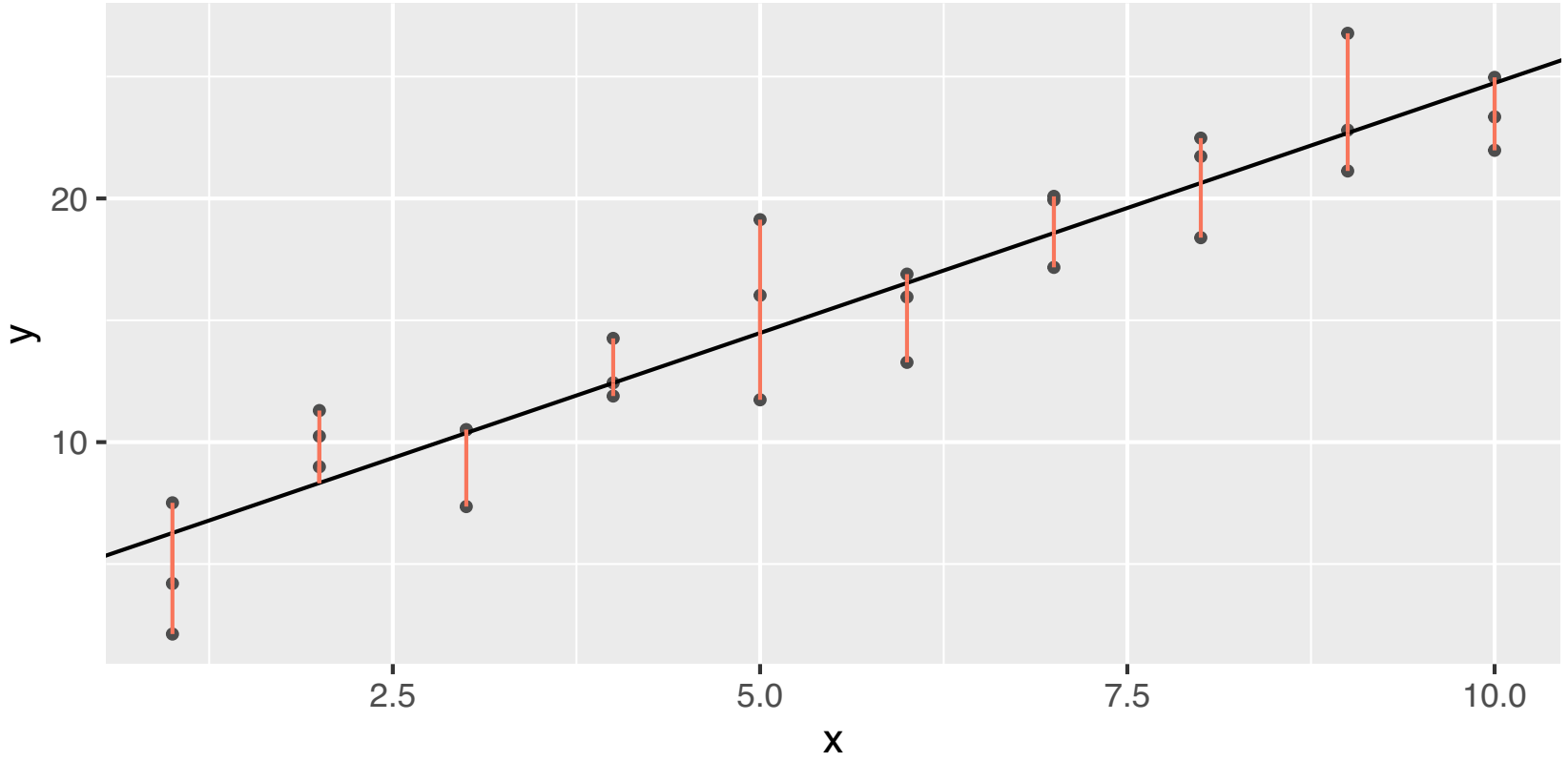
- Work with Simulated data set
  - package called "modelr"
- sim1 - data set that we will use.
- $\{(x_i, y_i) : 1 \le i \le n\}$ ≡ Data
- $f(a_1, a_2) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - (a_1 + a_2 x_i))^2}$
  - Found a_1, a_2 that minimized f
    - least square line

To do this entire procedure

R - in built function ≡ "lm"

Linear Models : best line

```
(Intercept)           x
   4.220822    2.051533
```



Using lm function

- check week 2 v code to get precise syntax.

Forbes data :-

~1857

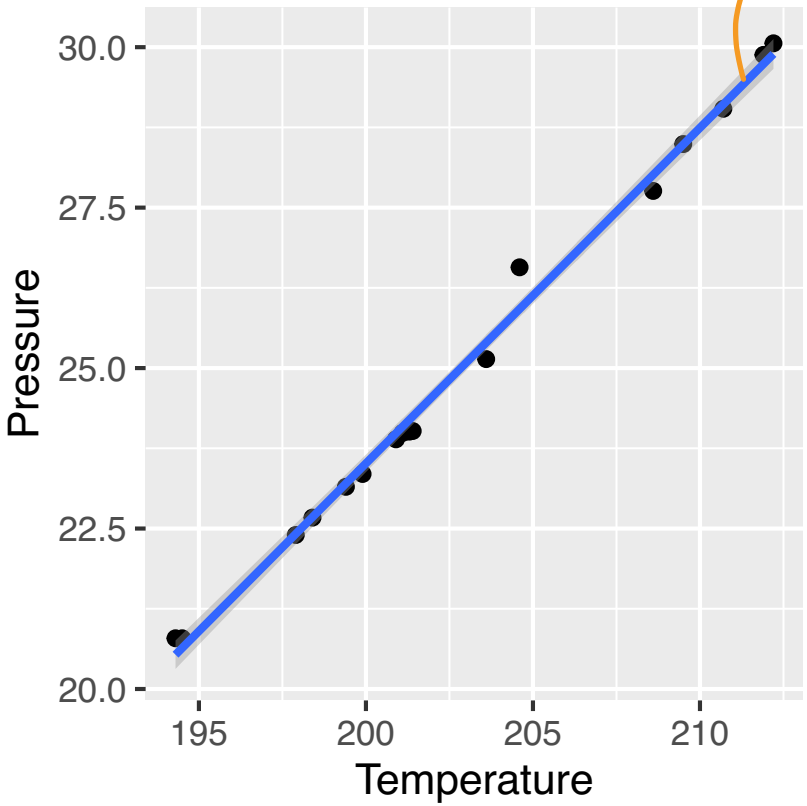
- atmospheric pressure & boiling point of water
- collected data: Alps & in Scotland

Could this
- replace the fragile Barometer

- 17 data points
- plot the data point
  - In function to find the best line

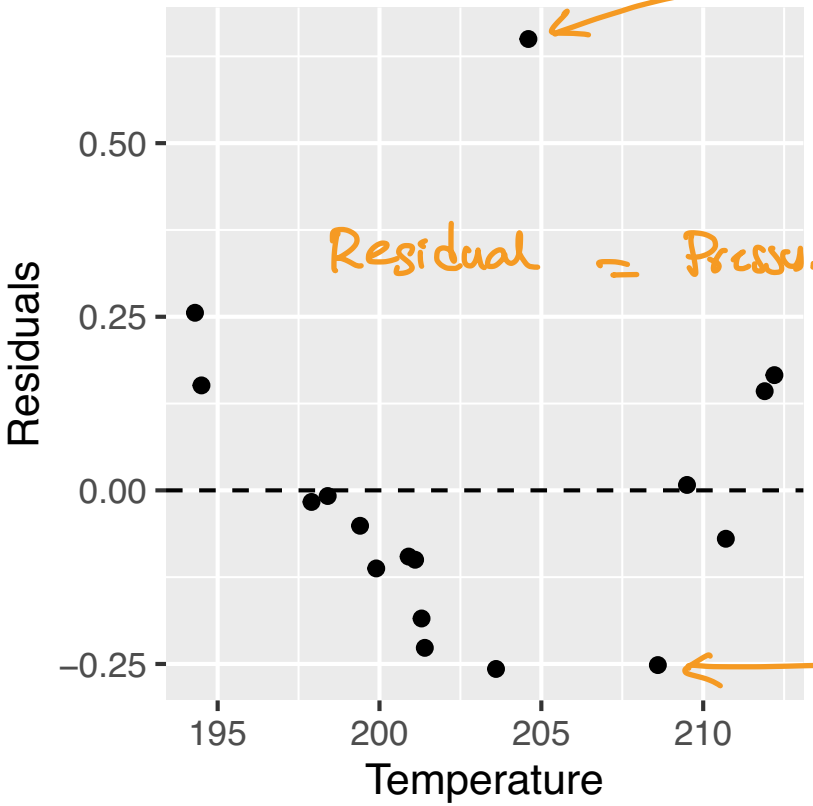
Forbes data

linear - best least square line



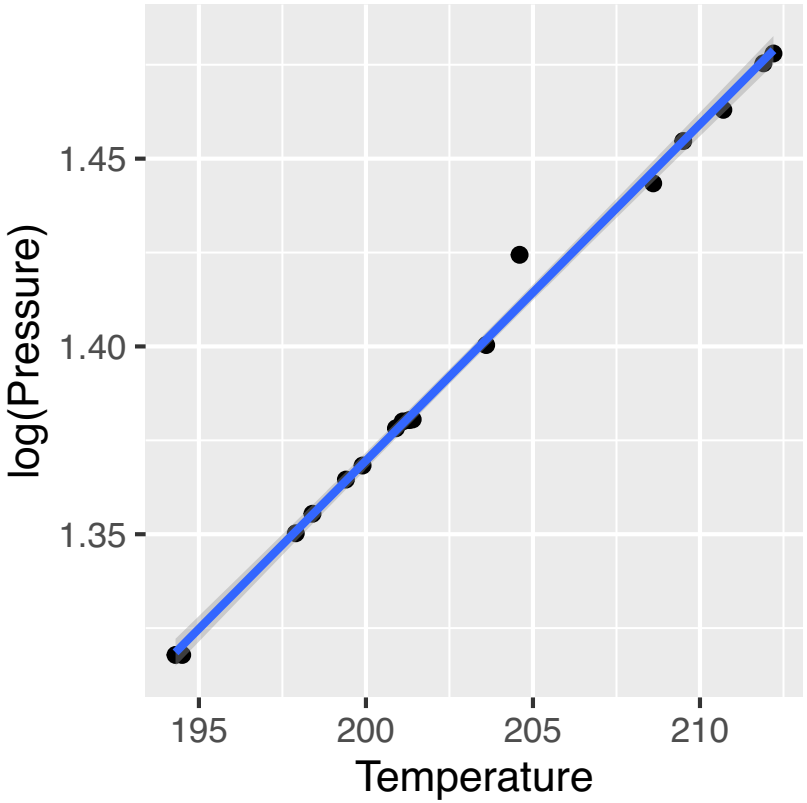
largest positive residual

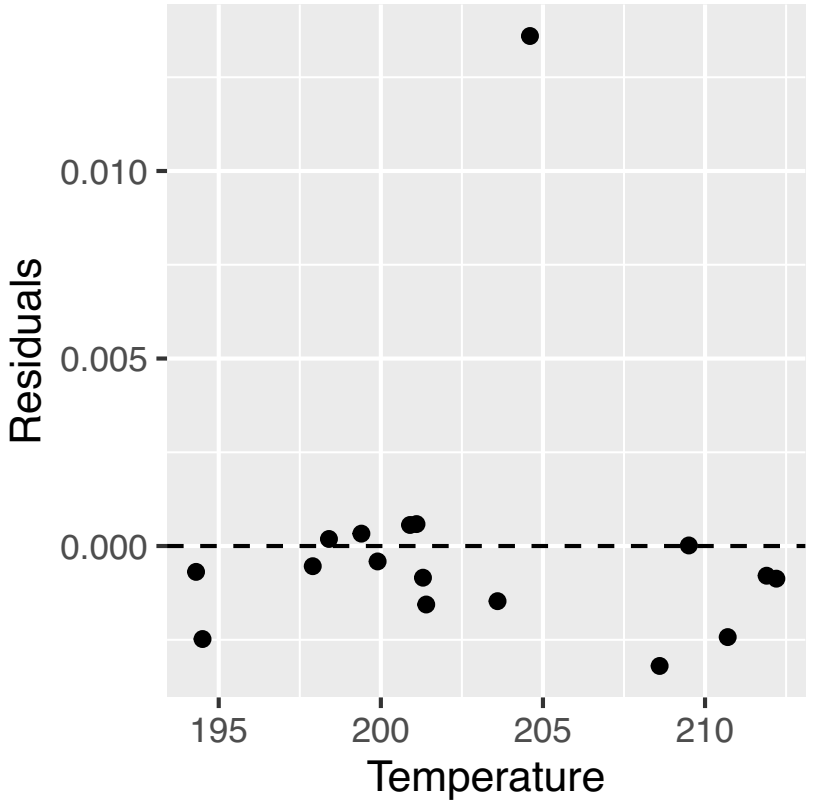
Residual = Pressure - "Point on line"



most negative & flat

Forbes: physical Theory ~ log(Pressure) = linearly related to temperature





- length (mm) at Age for smallmouth Bass (fish data set)
  - lake in Minnesota U.S.A ~ 1991

fish have annular ring like trees ← Age: ≤ 7 (use to compute)

- graph ≡ cannot use the line to predict age from length.

- Predicting the weather

Sept – Dec Snowfall ------> Predict Jan – June ?

93 years data

- early & late are not related!