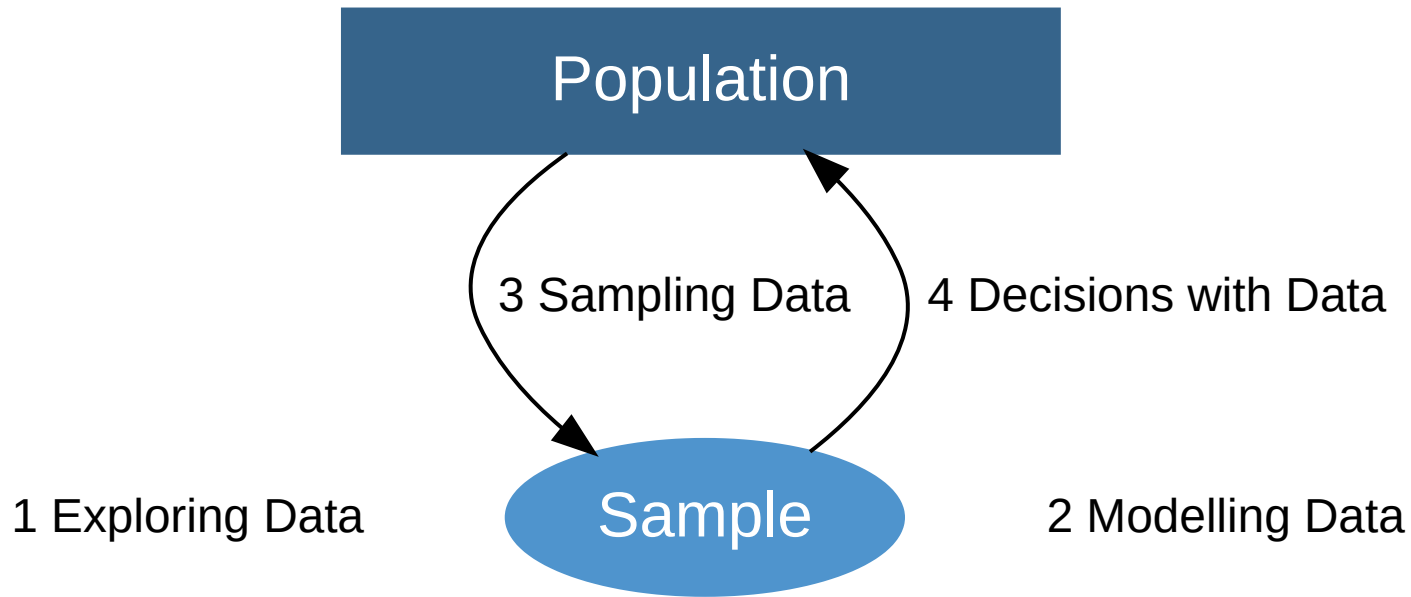


Regression Line

Modelling data | Linear Model

© University of Sydney DATA1001/1901

Unit Overview





Module2 Modelling Data

Normal Model

What is the Normal Curve? How can we use it to model data?

Linear Model

How can we describe the relationship between 2 variables? When is a linear model appropriate?



Regression Line

Data Story | How is the air quality in North-West Sydney related to Central-East Sydney?

Regression Line

Prediction

Summary

Data Story

How is the air quality in North-West Sydney related to Central-East Sydney?

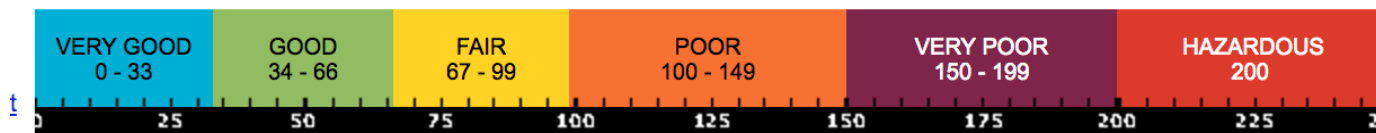
AQI data

- How do scientists monitor the air quality of Sydney?
- The [Office of Environment and Heritage](#) (OEH) has 14 active monitoring sites.



- At each site, data readings are taken for 6 pollutants:
 - Ozone
 - Nitrogen dioxide
 - Visibility
 - Carbon monoxide
 - Sulfur dioxide
 - Particles
- These are combined into the air quality index (AQI).

 Why?





Statistical Thinking

Who is the AQI index useful for?

- People with sensitive respiratory conditions (e.g. people with asthma, older adults and children) should consider either cutting back or rescheduling strenuous outdoor activities when air quality is 'poor' or worse.
- Environmental scientists studying changes in air quality.
- Potential home-buyers.

- We will consider the [data](#) for July 2015 for two regions:
 - Sydney's central-east (CE)
 - Sydney's north-west (NW)

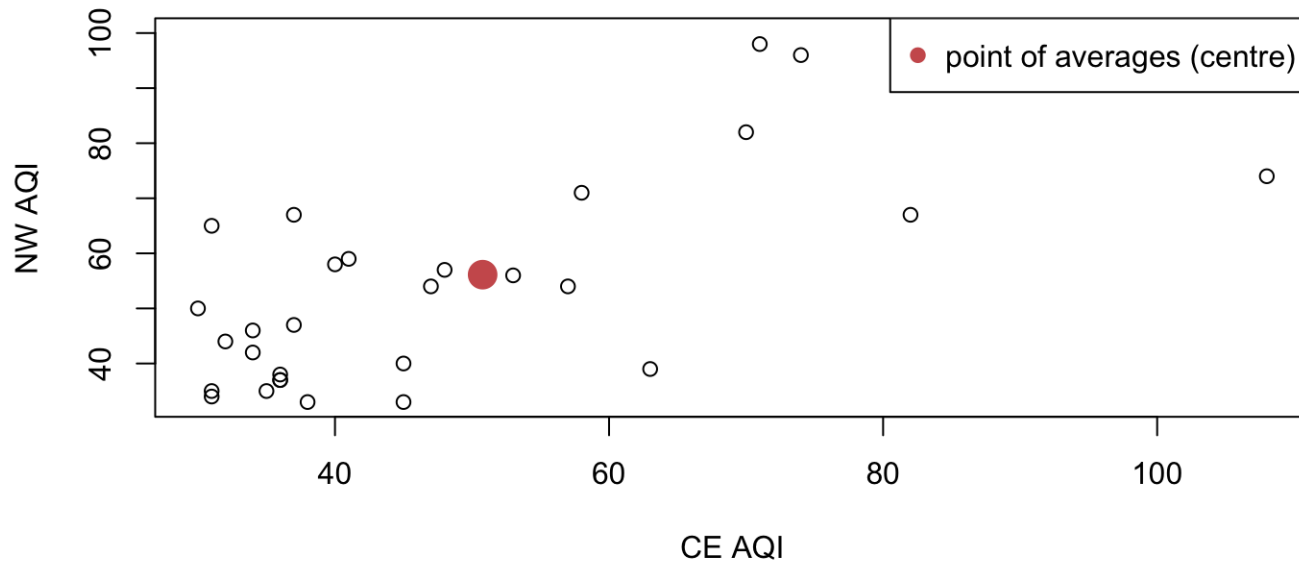
```
library(readxl)
data = read_excel("data/AQI_July2015.xls")
```

```
head(data)
```

```
## # A tibble: 6 x 3
##   Date      SydneyCEAQI SydneyNWAQI
##   <chr>      <dbl>      <dbl>
## 1 01/07/2015      99         92
## 2 02/07/2015      32         44
## 3 03/07/2015      70         82
## 4 04/07/2015      74         96
## 5 05/07/2015      95        100
## 6 06/07/2015      71         98
```

Scatter plot


```
CE = data$SydneyCEAQI
NW = data$SydneyNWAQI
plot(CE, NW, xlab="CE AQI", ylab="NW AQI")
points(mean(CE),mean(NW), col = "indianred",pch=19,cex = 2) # point of averages (centre)
legend("topright",c("point of averages (centre)"),col="indianred",pch=19)
```



Correlation Coefficient

```
cor(CE,Nw)
```

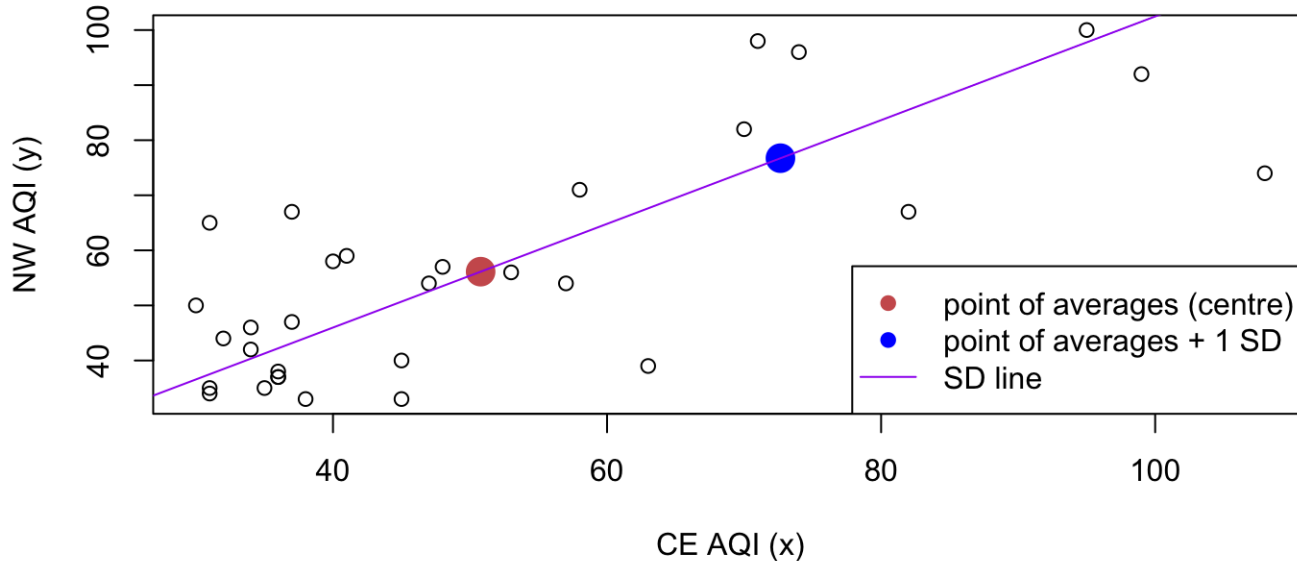
```
## [1] 0.757917
```

 What does the size of this correlation coefficient suggest about the data? How do we find the **optimal** line?

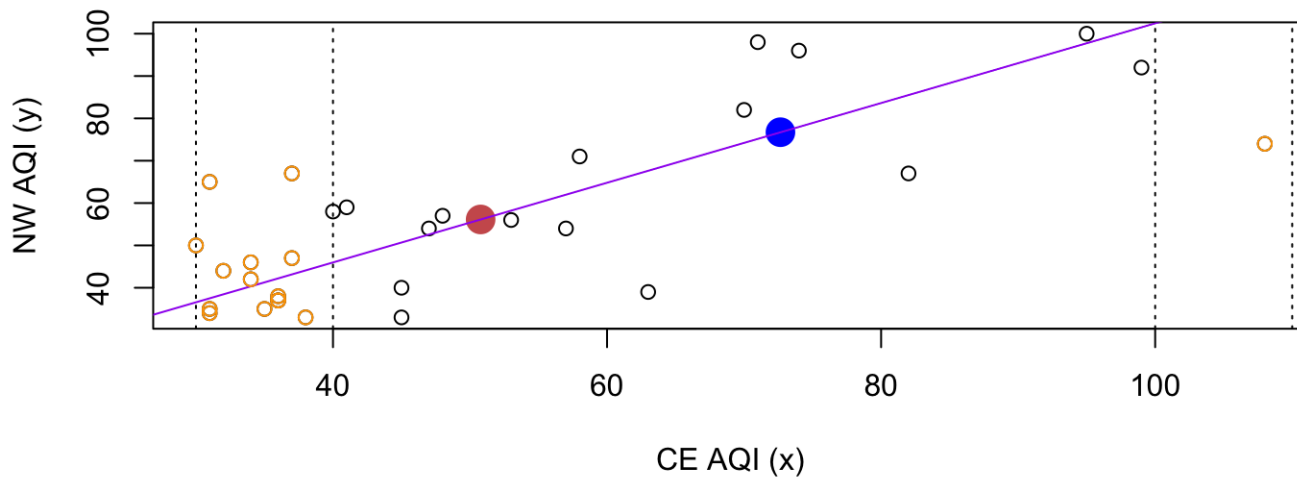
Regression Line

1st Option: SD Line

- Previously, the **SD line** visually looked like a good candidate as it connects the point of averages (\bar{x}, \bar{y}) to $(\bar{x} + \text{SD}_x, \bar{y} + \text{SD}_y)$ (for this data with positive correlation).

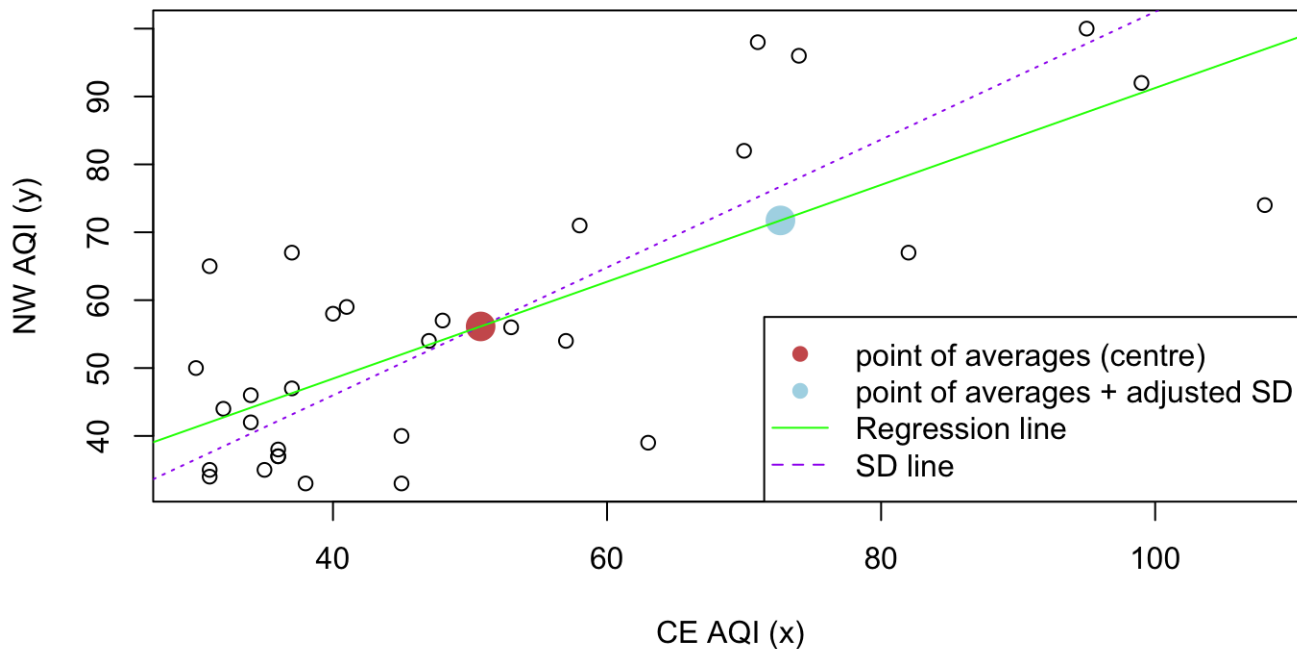


- However, it does not use the correlation coefficient, so it is insensitive to the amount of clustering around the line.
- Note how it underestimates (LHS) and overestimates (RHS) at the extremes.

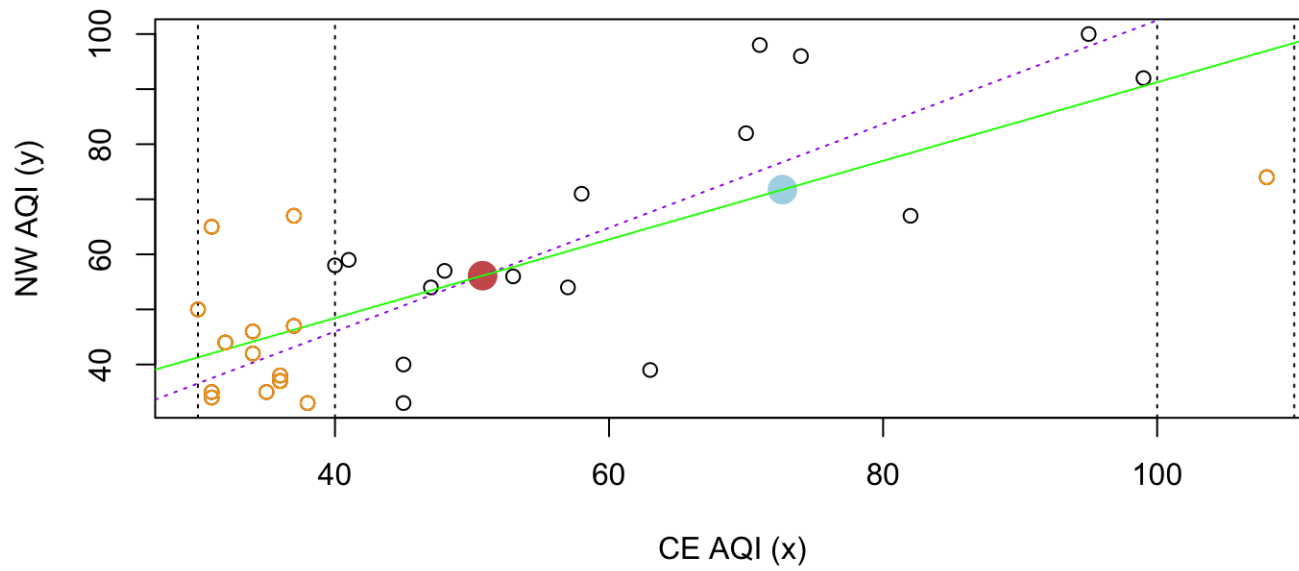


Best Option: Regression Line

- To describe the scatter plot, we need to use the 5 summaries: \bar{x} , \bar{y} , SD_x , SD_y , r .
- The **Regression line** connects (\bar{x}, \bar{y}) to $(\bar{x} + SD_x, \bar{y} + rSD_y)$



- Note the improvement at the extremes.



In R

```
lm(NW~CE)
```

```
##  
## Call:  
## lm(formula = NW ~ CE)  
##  
## Coefficients:  
## (Intercept)          CE  
##    19.8874      0.7138
```

```
model= lm(NW~CE)  
model$coeff
```

```
## (Intercept)          CE  
##  19.8873954    0.7137806
```

So for $x = \text{CE}$ and $y = \text{NW}$, the regression line is

$$y = 19.8874 + 0.7138x$$

Comparing Regression Line and SD Line (Ext)

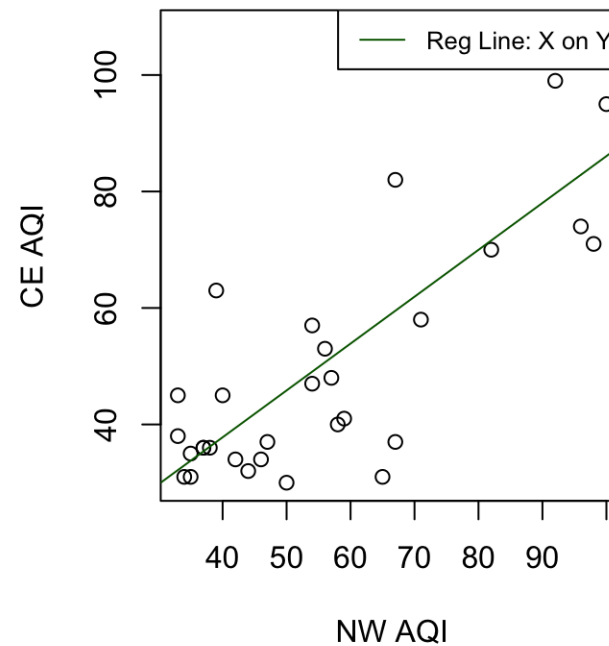
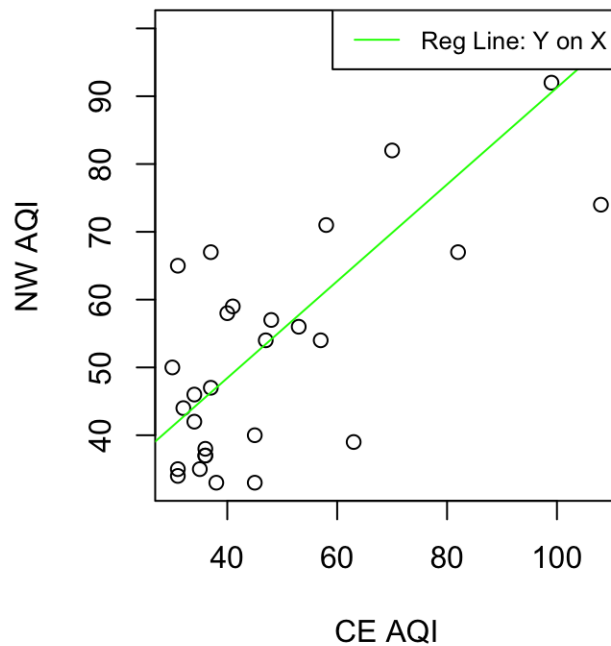
Formally, we could compare the 2 lines:

Feature	SD Line	Regression Line
Connects	(\bar{x}, \bar{y}) to $(\bar{x} + \text{SD}_x, \bar{y} + \text{SD}_y)$ ($r \geq 0$) (\bar{x}, \bar{y}) to $(\bar{x} + \text{SD}_x, \bar{y} + \text{SD}_y)$ ($r < 0$)	(\bar{x}, \bar{y}) to $(\bar{x} + \text{SD}_x, \bar{y} + r\text{SD}_y)$
Slope (b)	$\frac{\text{SD}_y}{\text{SD}_x}$ ($r \geq 0$) $\frac{-\text{SD}_y}{\text{SD}_x}$ ($r < 0$)	$r \frac{\text{SD}_y}{\text{SD}_x}$
Intercept (a)	$\bar{y} - b\bar{x}$	$\bar{y} - b\bar{x}$

We can derive the (least-squares) regression line using calculus (see the [Maths Guide](#)).

2 Regression lines

- We can predict Y from X or X from Y , depending on what fits the context.



Beware!

- We need to **refit** the model.
- We cannot simply rearrange the other equation, or interchange the intercept and the slope.

```
lm(CE~NW)
```

```
##  
## Call:  
## lm(formula = CE ~ NW)  
##  
## Coefficients:  
## (Intercept)          NW  
##      5.6025      0.8048
```

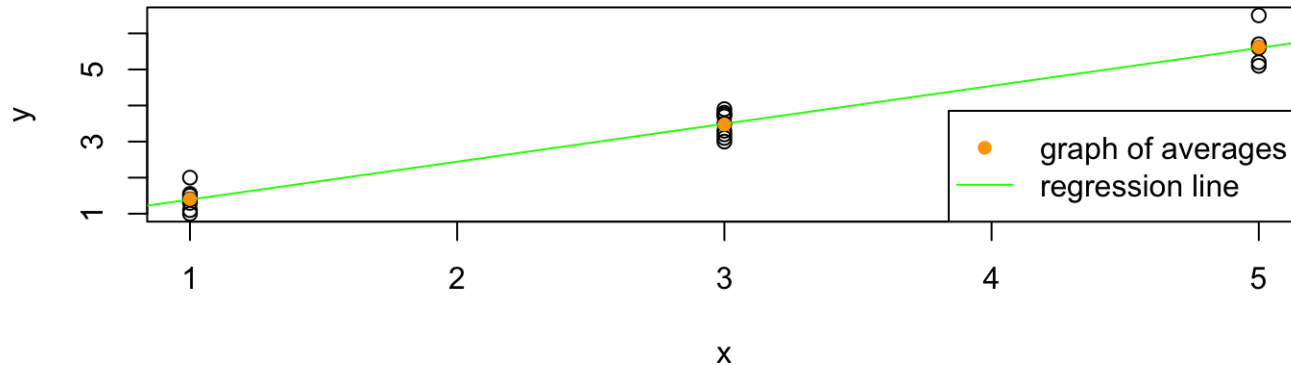
The graph of averages

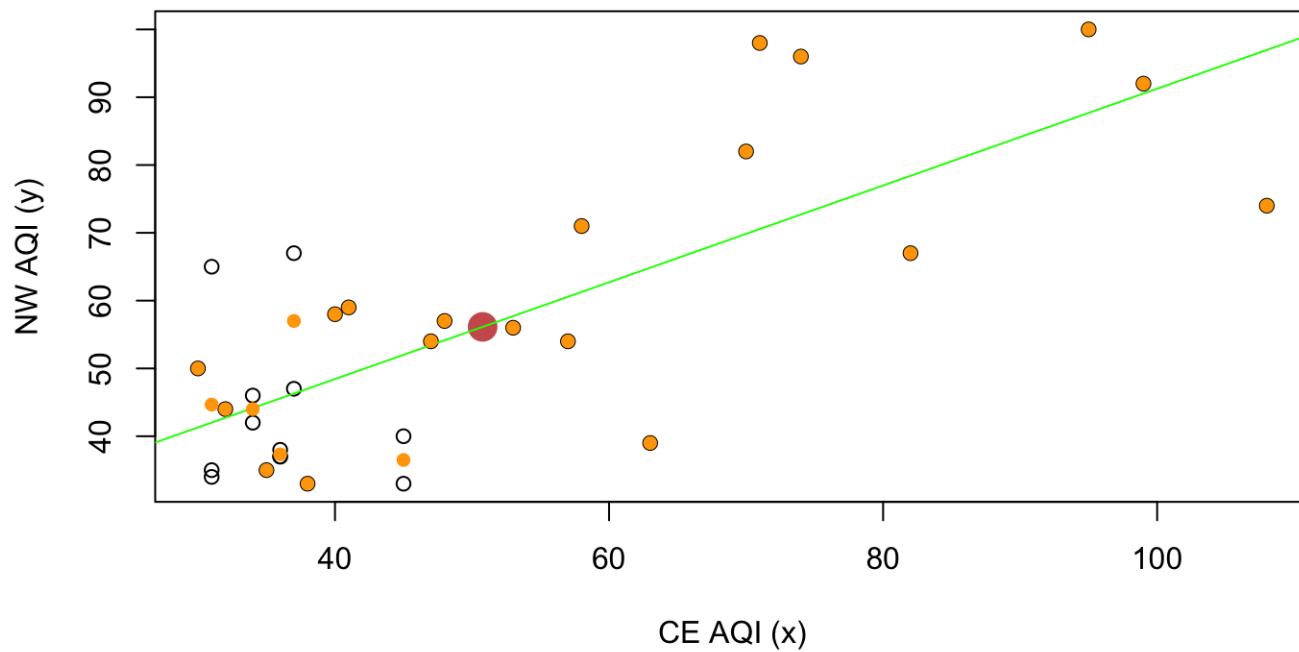


Graph of averages

The graph of averages plots the average y for each x .

- The regression line is a smoothed version of the graph of averages.
- If the graph of averages is a straight line, that line is the regression line.





Note here that there are not many repeats of the CE values, so most of the “averages” for NW are in fact the single data points.

Prediction

Need for predictions

- On a particular morning, the CE AQI is recorded, but the NW AQI reading is lost.
- How can we best predict the NW reading?



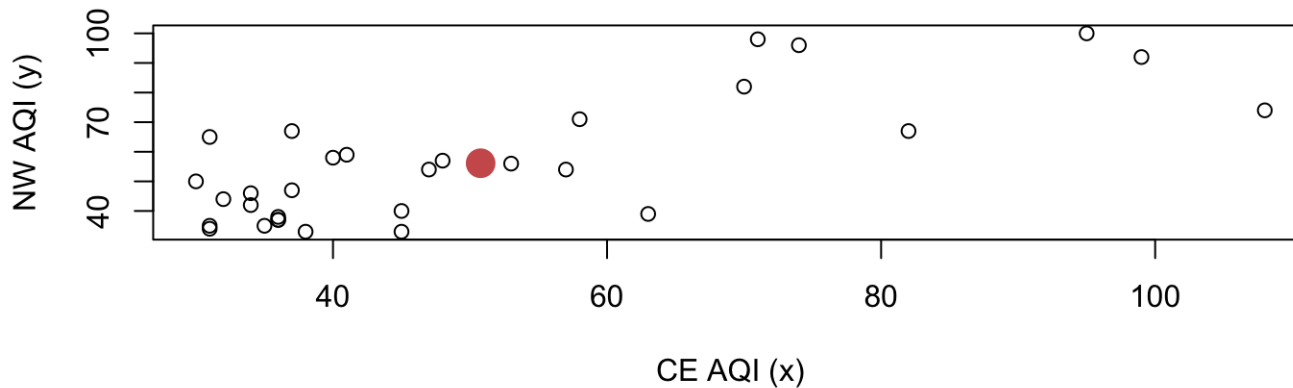
How can data be “lost”?

Method1: Baseline prediction

- Given a certain value x , a basic prediction of y would be the **average** of y over **all** the x values in the data.
- So for any CE reading, we could predict the NW air quality to be 56.13.

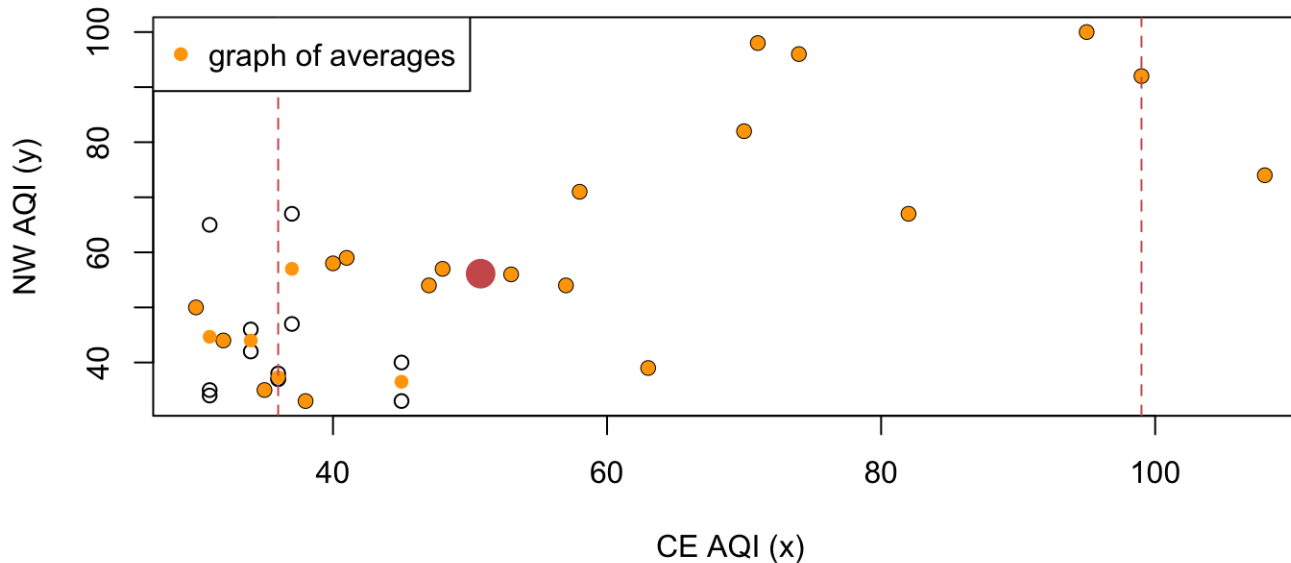
```
mean(NW)
```

```
## [1] 56.12903
```



Method2: Prediction in a strip

- Given a certain value x , a more careful prediction of y would be the average of all the y values in the data corresponding to that x value.
- We use the graph of averages.



Example1

For a CE reading of 99, we would predict the NW air quality to be 92.

```
data[CE==99,]
```

```
## # A tibble: 1 x 3
##   Date      SydneyCEAQI SydneyNWAQI
##   <chr>      <dbl>      <dbl>
## 1 01/07/2015      99          92
```

```
mean(NW[CE==99])
```

```
## [1] 92
```

Example2

For a CE reading of 36 (repeated 3 times), we would predict the NW air quality to be 37.3 (1dp).

```
data[CE==36,]
```

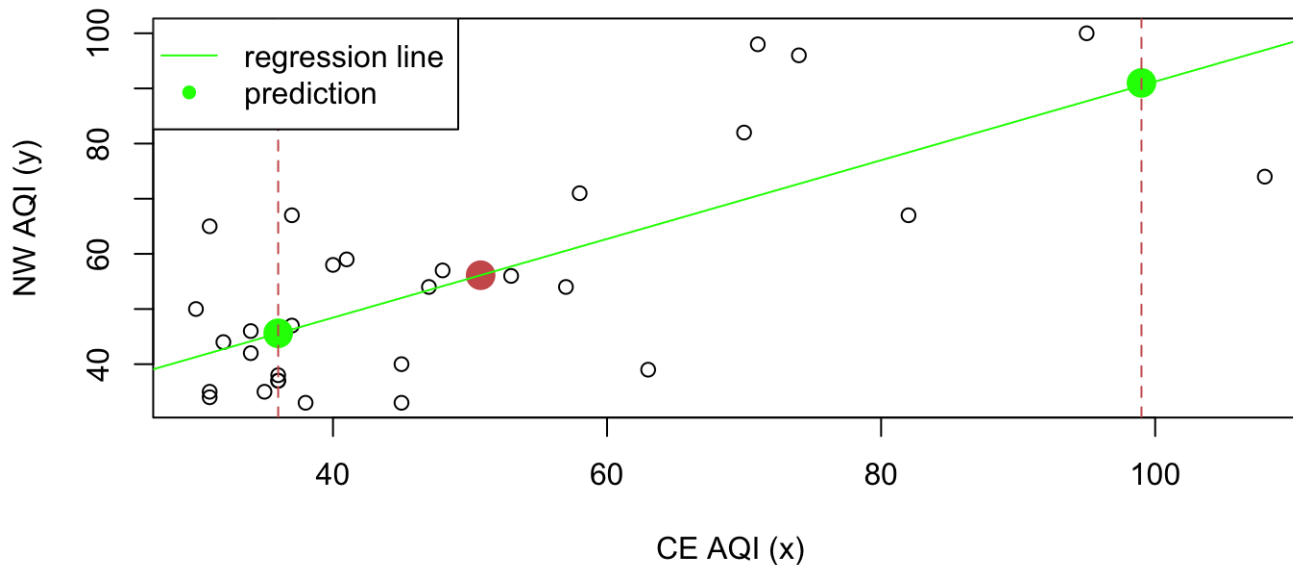
```
## # A tibble: 3 x 3
##   Date      SydneyCEAQI SydneyNWAQI
##   <chr>      <dbl>      <dbl>
## 1 25/07/2015      36          38
## 2 26/07/2015      36          37
## 3 27/07/2015      36          37
```

```
# Or subset(data, CE==36)
mean(NW[CE==36])
```

```
## [1] 37.33333
```

Method3: The Regression line

- The best prediction is based on the Regression line
- For AQI, we have $y = 19.8874 + 0.7138x$.



Example1

For a CE reading of 99, we would predict the NW air quality to be
 $y = 19.8874 + 0.7138 \times 99 \approx 91$ (1dp).

Example2

For a CE reading of 36, we would predict the NW air quality to be
 $y = 19.8874 + 0.7138 \times 36 \approx 45.6$ (1dp).

Method4: Predicting percentile ranks

If x is in a certain percentile of all the x 's, what percentile would we predict the corresponding y to be in?

Step1: Find the z score in the x direction : z_x .

Step2: Find the predicted z score in the y direction : $z_y = r * z_x$.

Step3: Translate z_y back to the percentile in the y direction.

Example1

When the CE reading is at the 90th percentile, we predict the NW reading to be at the 83th percentile.

```
z_x = qnorm(0.9)
z_y = cor(CE,NW)*z_x
pnorm(z_y)
```

```
## [1] 0.834303
```

Note: it would be mistake to assume that the 90% percentile of CE corresponds to the 90% percentile of NW, unless $r = 1$.

Example2

When the CE reading is at the 5th percentile, we predict the NW reading to be at the 11th percentile.

```
z_x = qnorm(0.05)  
z_y = cor(CE,NW)*z_x  
pnorm(z_y)
```

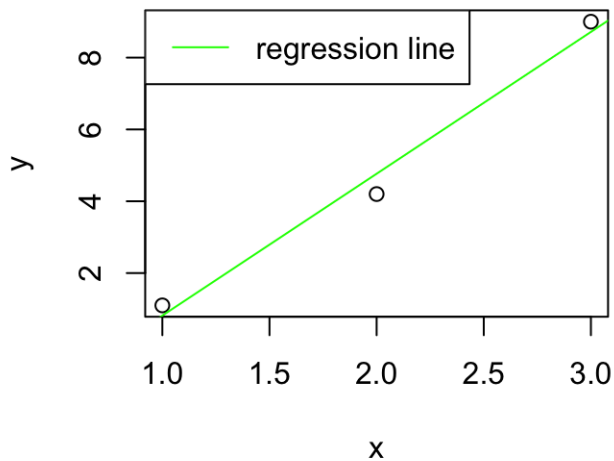
```
## [1] 0.1062606
```

Mistakes in Prediction

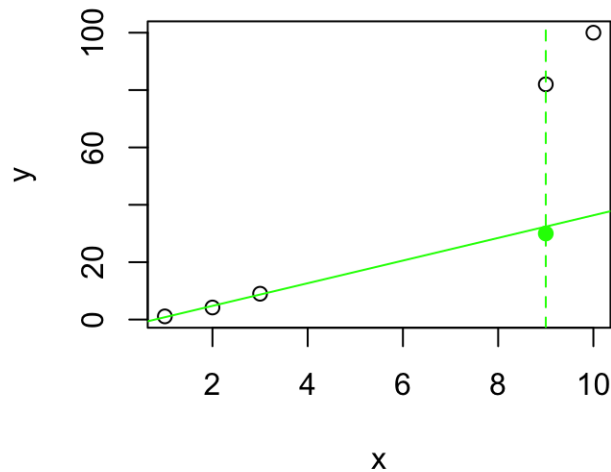
1. Extrapolating

If we make a prediction from an x value that is not within the range of the data, then that prediction can be completely **unreliable**.

Fitting line for 1st 3 data points

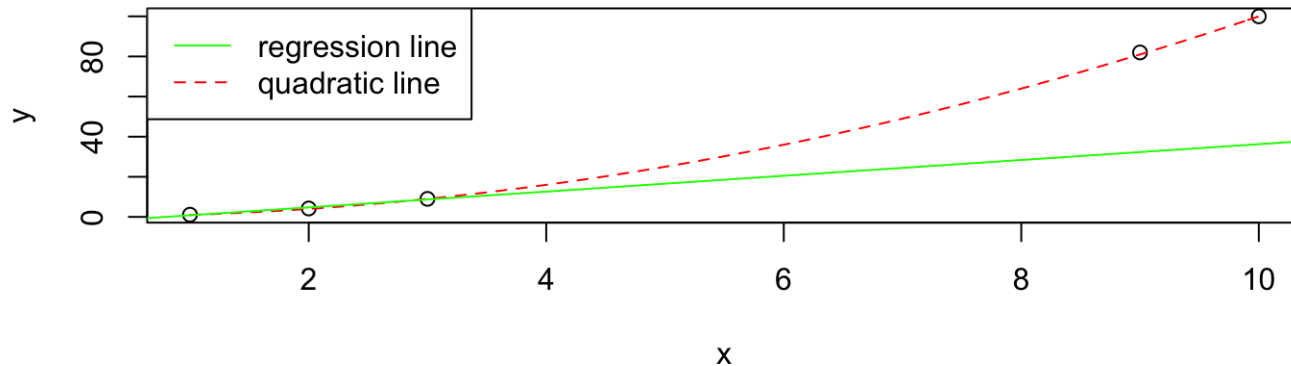


Long-term trend not linear



2. Not checking the scatter plot

- We can have a high correlation coefficient and then fit a regression line, but the data may not even be linear!
- So always check the scatter plot first!



Note: Even though the correlation coefficient is high $r = 1$, a quadratic model is more appropriate than a linear model.

- You can also check the **residual plot** which is a plot of the gaps between the actual values and the line. This shows up if any pattern has not been captured by fitting a linear model. If the linear model is appropriate, the residual plot should be a random scatter of points.

3. The regression fallacy (Ext)



The regression fallacy

- In virtually all **test-retest** situations:
 - the bottom group will improve on average
 - the top group will decline on average.
- Galton called this **regression to mediocrity** or the **regression effect**.
- The **regression fallacy** is being worried about the regression effect!

For a test-retest context, we would expect

$$\text{observed test score} = \text{true score} + \text{chance error}$$

- So if someone scores above average on the 1st test, then it is likely the chance error is positive.
- So the true score is lower than the observed score, which could be reflected in a lower retest score.

Summary

For prediction, the Regression line is better than the SD line as it uses all 5 numerical summaries for the scatter plot. It is a smoothed version of the graph of averages. The regression effect is observed in test-re-test situations, leading to the regression fallacy.

Key Words

SD Line, Regression Line, 2 regression lines, graph of averages, prediction, regression effect, regression fallacy