Week 5 5.2 Note

Lecture 13: Scatter Plot and Correlation

---

1. Bivariate Data

> **Bivariate data** involves a *pair* of variables.

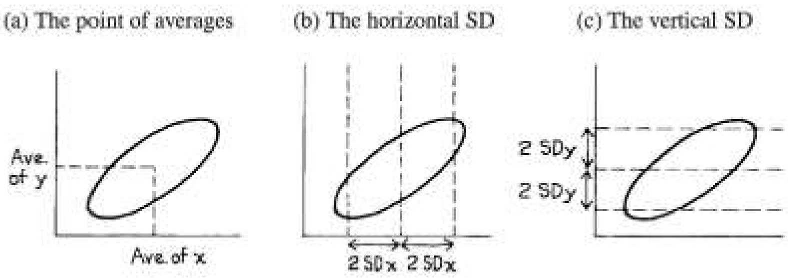
2. Linear Association

> The **linear association** between *2 variables* describes how tightly the points *cluster* around a line.

- If one variable tends to *increase* with the other, then we have *positive* association.

3. Correlation Coefficient

- How can we summarize a scatter plot?
    - Mean and SD of X (x, SDx)
    - Mean and SD of Y (y, SDy)
    - Correlation Coefficient (r)
- Centre and Spread of the Cloud
    - The **centre** of the cloud is represented by the point of averages (x, y) (x and y here are means).
    - The **horizontal spread** of the cloud is measured by *SDx*, we expect most of the points to fall with *2 SDs* from x.
    - The **vertical spread** of the cloud is measured by *SDy*, we expect most of the points to fall with *2 SDs* from y.



> **Correlation Coefficient** (r): A numerical summary which measures the *clustering* around the line.

- It indicates the sign of strength of the linear association.
- Range: -1 to 1
    - If r is positive: the cloud slopes up.
    - If r is negative: the cloud slopes down.
    - More closer to +1 or -1: the points cluster more tightly.
- The population CC (rpop) is *the mean of the product of the variables* in **standard units**.
- use `cor()` for CC calculation. ex. `cor (data$fheight, data$sheight)`

4. SD Line

> **SD line** is the line that the points *cluster around*.

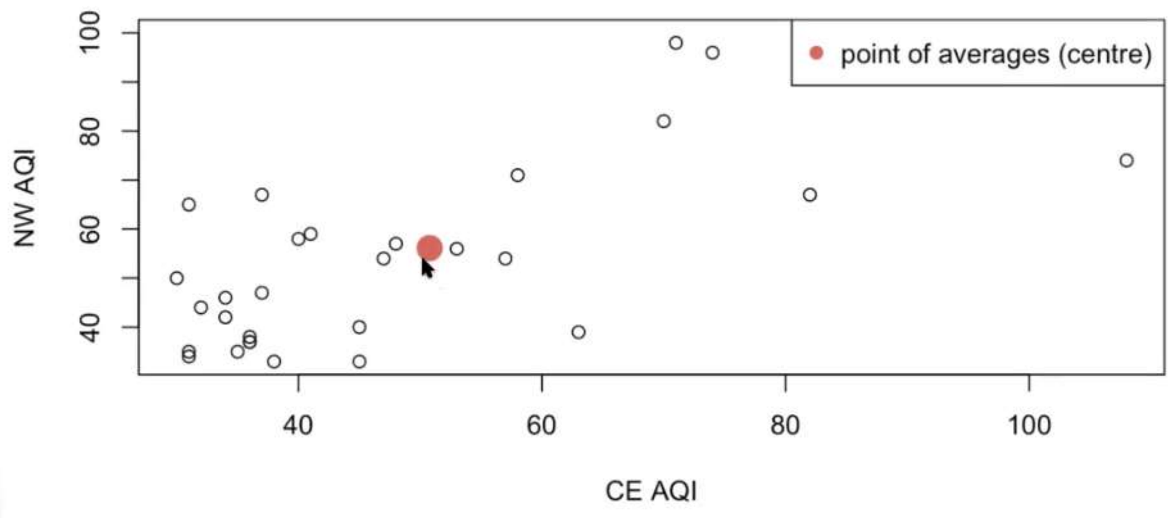
- It connects the point of averages (x, y) to (x+SDx, y+SDy for r > 0) or (x, y) to (x+SDx, y-SDy for r < 0)

Lecture 14: Scatter Plot and Correlation

---

1. Scatter Plot Example

```
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)
```



2. Properties of Correlation Coefficient

- When r=+1 or -1, all the points lie on a line (no cloud, perfect correlation)
- When r=0, the points don't fit around a line.
- CC is **scale invariant** (won't change)

3. Misleading Correlations

- *Outliers* can overly influence the CC.
  - Example:

```
# Add one outlier respectively to both data sets:
CE1 = c(CE, 100)
NW1 = c(NW, 20)
# Calculate the CC of the original:
cor(CE, NW)
0.757917
# Calculate the CC of the new:
cor(CE1, NW1)
0.5575432
# The CC has changed a lot by outliers.
```

- Nonlinear association can't be detected by the CC.
- The same CC can arise from different data.
  - Example: [Anscombes Quartet]
- Rates of averages tend to inflate the CC.

> Ecological correlation (spatial correlation): the correlation between 2 variables that are group means or rates.

  - EC tend to overestimate the association between 2 variables.
- Small SDs can make the correlation *look* bigger.

Lecture 15: Regression Line

---

1. Regression Line

- To describe the scatter plot, we need to use the 5 summaries: x, y, SDx, SDy, r.
- The regression line connects (x, y) to (x+SDx, y+SDy)
- Command: `lm()` ex. `lm(NW~CE)`

Formally, we could compare the 2 lines:

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

2. The Graph of Averages

- Plot the average y for each x.
- If the GOA is a *straight* line, that line is the *regression line*.

3. Prediction

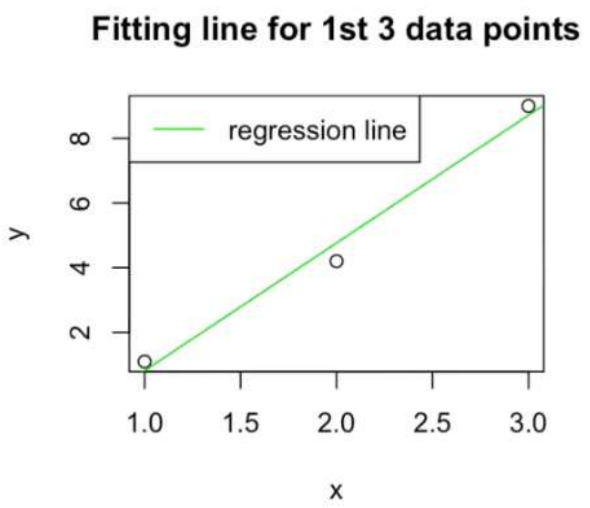
- Baseline Prediction:

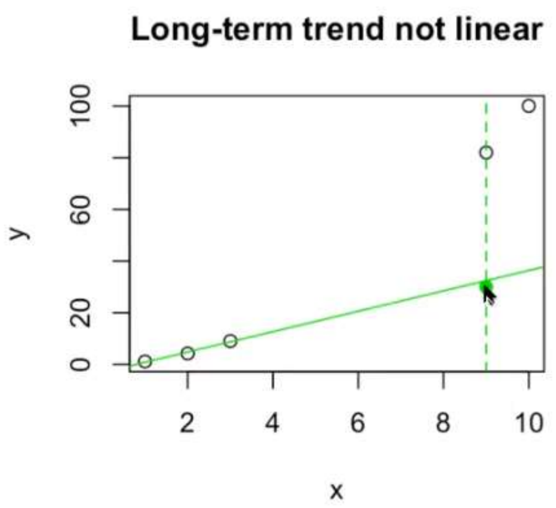
- Give a *certain value x*, a *basic* prediction of y would be **the average of y over all the x values**.
- Prediction in a strip:
  - Give a *certain value x*, a *more careful* prediction of y would be **the average of all the y values** in the data corresponding to **that x value**.
- The Regression Line
  - Calculate the regression line, and insert the particular x value we will use.
- 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?
  - Steps:
    1. Find the z score in the x direction: Zx.
    2. Find the predicted z score in the y direction: Zy=r*Zx
    3. Translate Zy back to the percentile in the y direction.
  - Example:

```
# Find the percentile using qnorm(), in this example the CE reading is at 90th percentile (90%):
z_x = qnorm(0, 9)
# Scale the CC and Zx:
z_y = cor(CE, NW) * z_x
# Translate to the y percentile using pnorm():
pnorm(z_y)
0.834303
# Conclusion: When CE reading is at the 90th percentile, the NW reading is at the 83th percentile.
```

4. Mistakes in Prediction

- Extrapolating
  - If we make a prediction from x that is *not within the data range*, the prediction can be unreliable.





- Not Checking the Scatter Plot
  - We can have a high CC and then fit a regression line, but the data may not be linear (more appropriate to use a quadratic model).

Lecture 16: Residual Plot

1. Residual

> Residual (prediction error): the **vertical distance** (gap) of a point above and below *the regression line*.

- It represents the error between the actual value and the prediction: ei=yi-^yi, wherew yi is the actual value and ^yi is the prediction.
  - Example:

```
# Find the regression line:
l = lm(NW~CE)
# Use the 10th reading in NW data to minus the prediction (fitted.values):
NW[10] - l$fitted.values[10]
10
-11.41741
```

  - Or more quickly:

```
l$residuals[10]
10
-11.41741
```

2. The RMS Error

- Represent the **average gap** between the points and the regression line.

$$\text{RMS error}_{pop} = \text{RMS of (gaps from the line)} = \sqrt{\text{mean of (gaps)}^2}$$

More formally, $\text{RMS error}_{pop} = \sqrt{\frac{e_1^2+e_2^2+\dots e_n^2}{n}}$.

- Example:

```
res = NW - l$fitted.values
sqrt(mean(res^2))
13.22338
```

- For Baseline Prediction:
  - The RMS error is **the SD** for y.
- Speedy Way for Population RMS Error:

$$\text{RMS error}_{pop} = \sqrt{1 - r^2}SD_y$$

- Example:

```
# For sample:
sqrt(1-(cor(CE, NW))^2)*sd(NW)
13.44196
# For population:
sqrt(1-(cor(CE, NW))^2)*sd(NW)*sqrt((length(CE)-1)/(length(CE)))
13.22338
```

- Special Cases
  - Perfect Correlation: r = +1 or -1
    - RMS Error = 0, as all points lie on the line.
  - r = 0
    - RMS Error = 0, as the regression line is no help in predicting y.
  - Smallest RMS Error
    - The smallest is for the regression line.

3. Residual Plot

- Graphs the residuals vs x.
- If the linear fit is appropriate for the data, it should show no pattern.
- Command: `lm$Residuals`
  - Example:

```
l = lm(NW~CE)
plot(CE, l$residuals, ylab="residuals")
```

4. Vertical Strips

- If the VSs on the scatter plot show *equal spread* in the *y* direction, then the data is **homoscedastic**.
  - The RMS error can be used as a measure of spread for individual strips.
- If not, then the data is **heteroscedastic**.
  - The RMS error cannot.
- If *homoscedastic*, then we can use the **normal approximation** within the VSs.
  - We consider the y within the strip as *y** with:

$$\bar{y}^* = \bar{y} + z_x r SD_y$$

$$SD_y^* \approx \text{RMS Error}$$

where z_x is the z-score for the strip.

  - Example:

```
# Percentage of days that CE is above 90:
length(which((CE>90)))/length(CE)
0.09677419
# Calculate the normal approximation:
z=(90-mean(CE)/sd(CE))
1-pnorm(z)
0.0365018
```

Lecture 17: Linear Regression Summary (given bivariate data)

- Steps:

1. Produce a scatter plot----does it look *linear*?
2. Produce a regression line: y = a + bx
3. Calculate the CC (r)----how strong is the *linear regression*?
4. Produce a residual plot----does it look *random*? Is linear model good?
5. Check assumptions----does the data look *homoscedastic*?
6. Perform predictions----predict y for given x and y within a VS