Rensselaer

why not change the world?®

Lab exercises: beginning to work with data: distributions, correlations, Linear Regression, visualization exercises using ggplot2 package

Ahmed Eleish

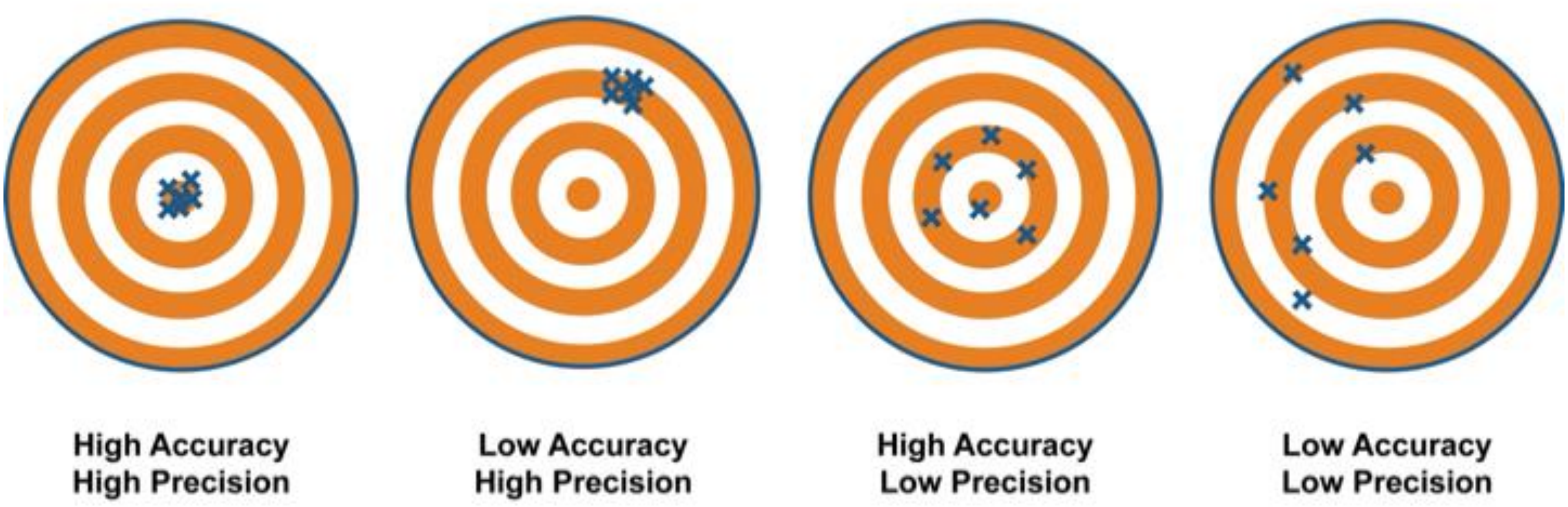
ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960 Group 1, Lab 1,

September 19th, 2024

Tetherless World Constellation
Rensselaer Polytechnic Institute

Lab 01 review

Accurate vs. Precise



http://climatica.org.uk/climate-science-information/uncertainty

Quantile-Quantile (Q-Q) Plot

- qqplot() function produces a quantile-quantile (Q-Q) plot.

- A *quantile-quantile (Q-Q) plot*, also called a *probability plot*, is a plot of the observed order statistics from a random sample (the empirical quantiles) against their (estimated) mean or median values based on an assumed distribution, or against the empirical quantiles of another set of data (Wilk and Gnanadesikan, 1968).

- **Q-Q plots are used to assess whether data come from a particular distribution, or whether two datasets have the same parent distribution**.
- **If the distributions have the same shape (but not necessarily the same location or scale parameters), then the plot will fall roughly on a straight line.**
- **If the distributions are exactly the same, then the plot will fall roughly on the straight line y=x**.

Lab 02

Files:

https://rpi.box.com/s/dyuhis8k5m1qyf58emakb8yyqmjwrnx7

Exercise 1: fitting a distribution beyond histograms

- Quantile-Quantile

```
help("qqnorm") #read the Rstudio documentation for qqnorm
qqnorm(EPI.new); qqline(EPI.new)
```

- Make a Q-Q plot against the generating distribution by:

```
x <- seq(20., 80., 1.0)
qqplot(qnorm(ppoints(200)), x)
qqline(x)
qqplot(qnorm(ppoints(200)),EPI.new)
qqline(EPI.new)
```

- Cumulative density function

```
plot(ecdf(EPI.new), do.points=FALSE)
plot(ecdf(rnorm(1000, 45, 10)), do.points=FALSE) # ecdf of normal distr with mean=45, sd= 10
lines(ecdf(EPI.new))
```

Exercise 1: fitting a distribution

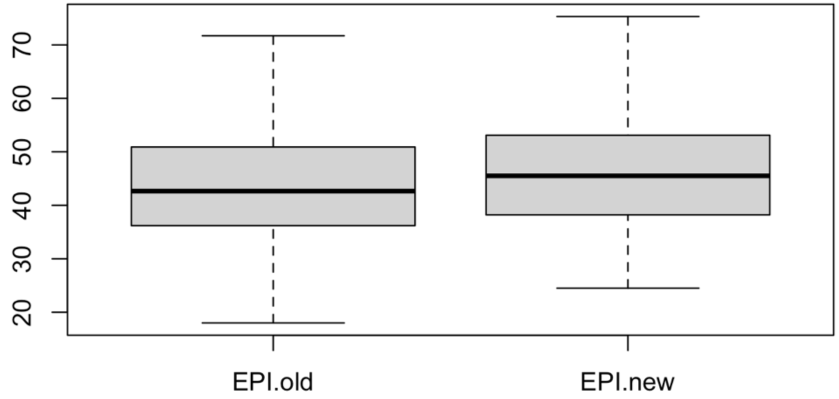
- Your exercise: do the same exploration and fitting for another 2 variables in epi2024results06022024.csv, i.e. primary variables (BDH, ECS, ...)

- You can find titles of these abbreviated indexes in epi2024weights.csv

- Try fitting other distributions – i.e. as ecdf or qq-plot e.g. qchisq, qbeta, qweibull

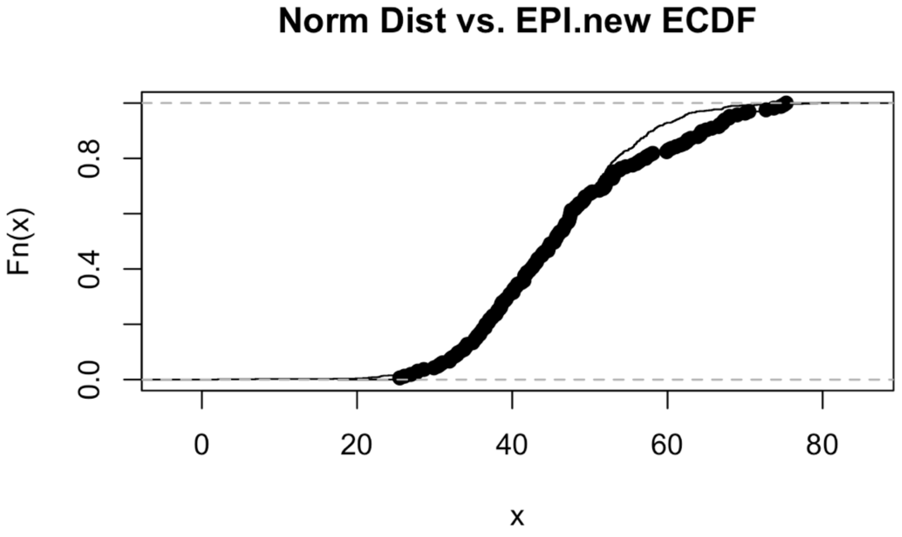
Comparing distributions

```
boxplot(EPI.old, EPI.new, names=c("EPI.old","EPI.new"))
```



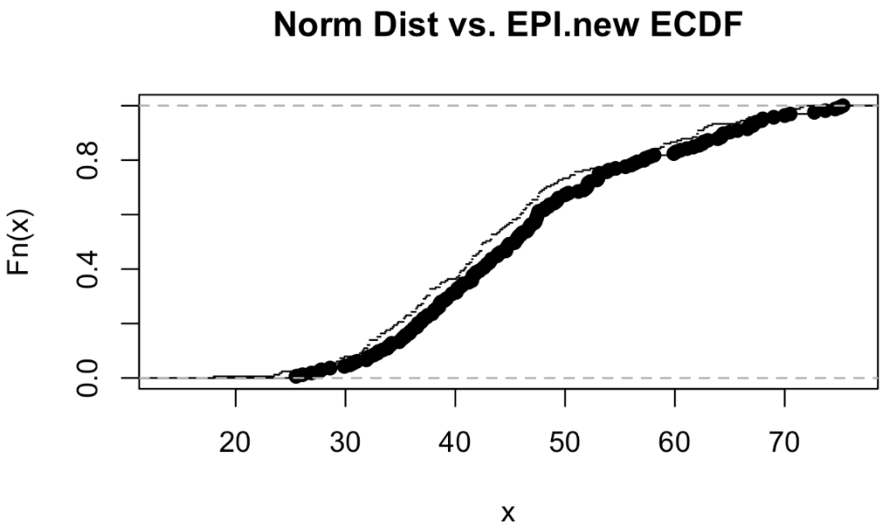
Comparing distributions

```
plot(ecdf(rnorm(1000, 45, 10)), do.points=FALSE)
lines(ecdf(EPI.new))
```



Comparing distributions

```
plot(ecdf(EPI.old), do.points=FALSE, main="EPI.old vs. EPI.new ECDF")
lines(ecdf(EPI.new))
```

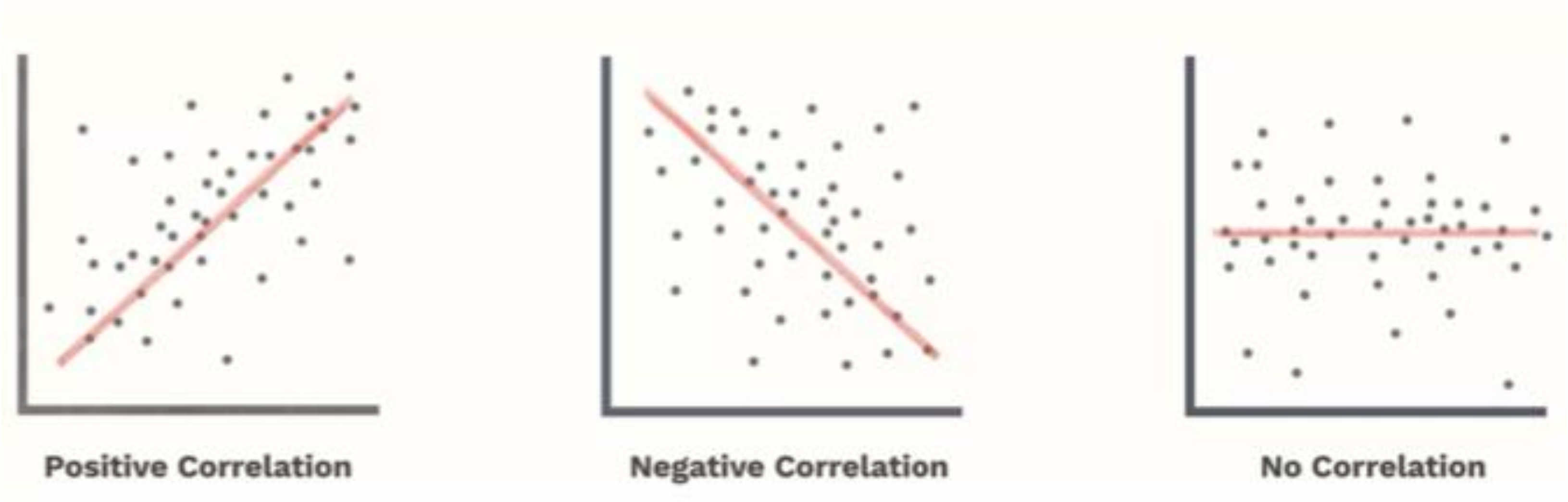


Correlation

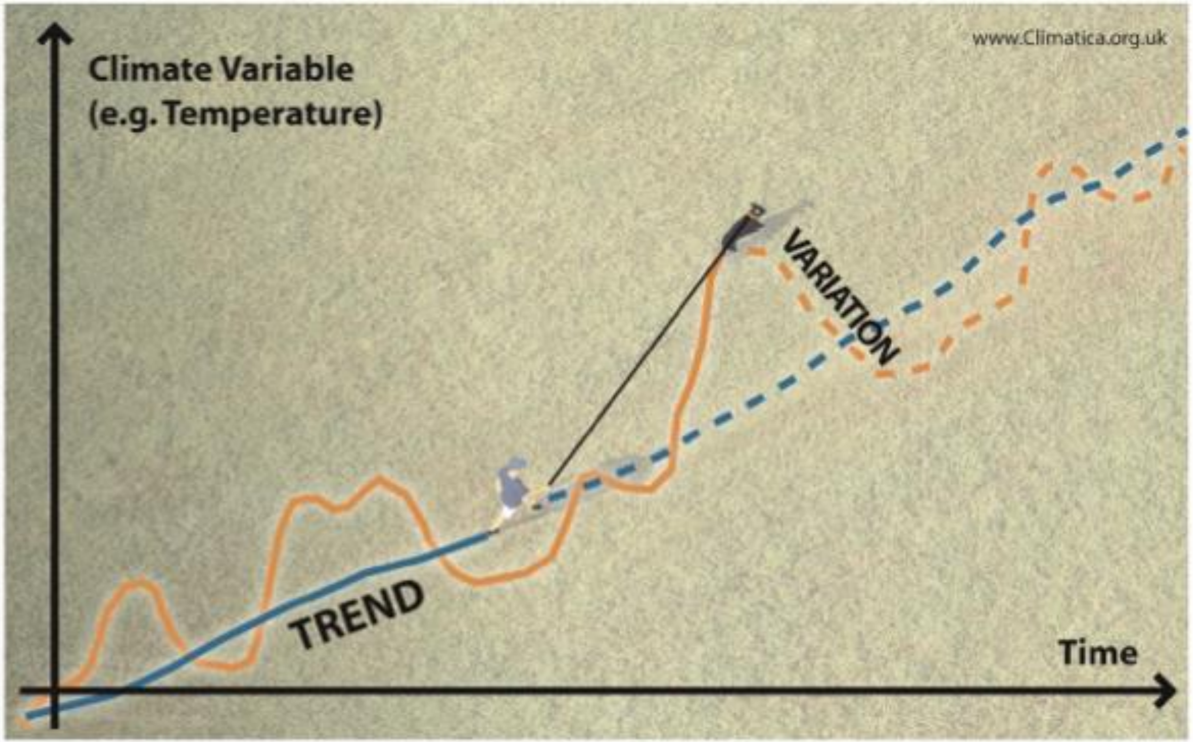
- One measure of the strength of the association between two numerical variables is correlation.
- Correlation describes the strength of the linear association between two variables.
- Correlation coefficient is between -1 and +1
  - -1 indicates a perfect negative linear association and +1 indicates a perfect positive linear association. The correlation coefficient of 0, indicates that there is no linear relationship in the two variables. -0.1 and +0.1, indicates no linear relationship *or a very weak* linear relationship
- Correlation coefficient is sensitive to outliers.
- Correlation coefficient is unitless.

Reference(s): https://www.investopedia.com/terms/c/correlationcoefficient.asp
https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp

Correlation...



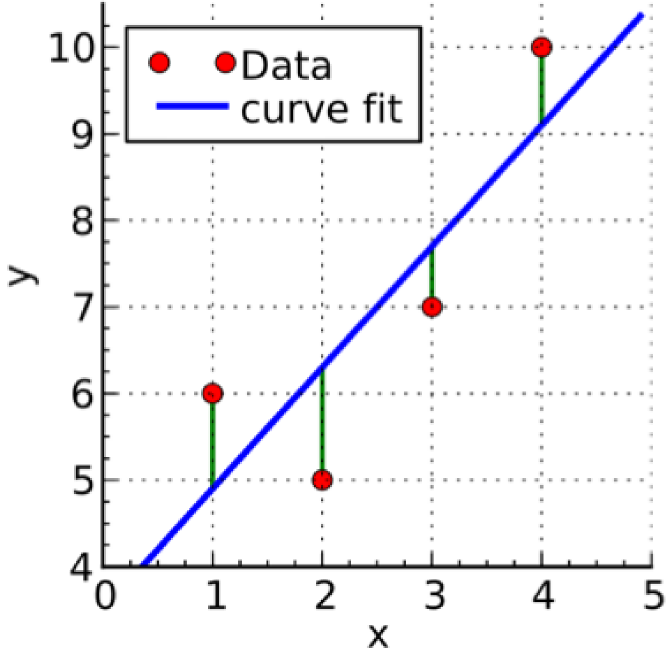
Image/Photo Credit: https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp



Person walking a dog. The dog's path may vary as it wanders from a straight line, but both the dog and the person are headed the same way. You can predict where they will end up by looking at the trend of the person walking the dog.

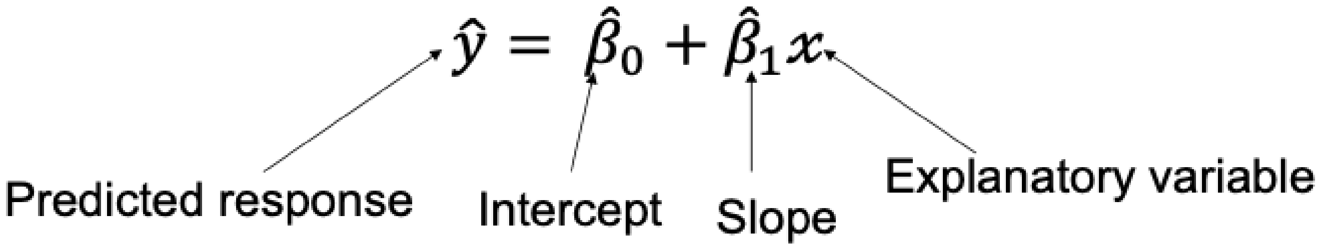
http://climatica.org.uk/climate-science-information/uncertainty

Regression



Simple Linear Regression

- Most commonly used approach is the *Least Squares*
- Least Squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS



- $\hat{y}$ = Predicted value of the response variable
- x = Explanatory variable (x)
- $\hat{\beta}_0$ = Intercept
- $\hat{\beta}_1$= Slope

Residuals ...

- The residual is defined as the difference between the observed value and the predicted value.(Difference between the observed value and the predicted value of the response variable for a given data point).

$e_i = y_i - \hat{y}_i$ represents the i_{th} residual,

this is the difference between the i_{th} observed response value and the i_{th} response value that is predicted by the linear model.

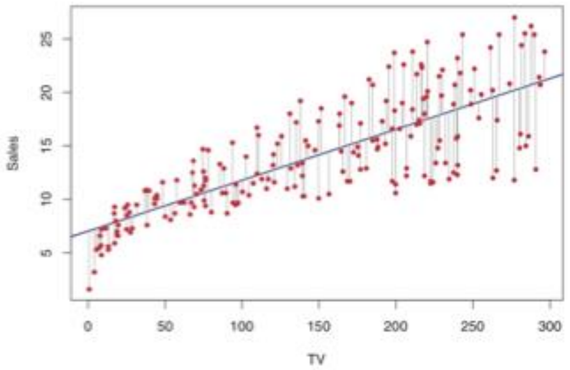


Image Credit: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 3 – Linear Regression

Reference: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 3 - Linear Regression

- How do we measure the best line?
- There are two options:

Option 1: Minimize the sum of magnitudes (absolute values) of the residuals

$$|e_1| + |e_2| + |e_3| + \ldots + |e_n|$$

OR

Option 2: Minimize the sum of squared residuals

$$\text{RSS} = e_1 + e_2 + e_3 + \ldots + e_n$$

$e_i = y_i - \hat{y}_i$ represents the *ith* residual

Reference: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 3 - Linear Regression

Outliers in Regression

- How does an outlier influence the least square line?
- In general, outliers are the points that fall away from the cloud of points.
- Two types –

– Leverage Points : Outliers that fall horizontally away from the center of the cloud of points but don't influence the slope of the regression line are called leverage points.

– Influential Points : Outliers that actually influence the slope of the regression line are called influential points.

Exercise 2: linear models

- Your exercise: Examine the influence of population on various indexes (EPI, ECO, etc.)

- You can find the populations in countries_populations_2023.csv

To integrate datasets:

```
# read data
populations_2023 <- read.csv("~/Courses/Data Analytics/Fall24/labs/lab01/countries_populations_2023.csv")

# drop countries not in epi results
populations <- populations_2023[-which(!populations_2023$Country %in% epi.results$country),]

# sort populations by country
populations <- populations[order(populations$Country),]

# drop countries not in populations
epi.results.sub <- epi.results[-which(!epi.results$country %in% populations$Country),]

# sort epi results by country
epi.results.sub <- epi.results.sub[order(epi.results.sub$country),]

# only keep necessary columns
epi.results.sub <- epi.results.sub[,c("country","EPI.old","EPI.new")]

# convert population to numeric
epi.results.sub$population <- as.numeric(populations$Population)

# compute population log base 10
epi.results.sub$population_log <- log10(epi.results.sub$population)
```

Linear Model in R

```
attach(epi.results.sub)

lin.mod.epinew <- lm(EPI.new~population_log,epi.results.sub)

plot(EPI.new~population_log)
abline(lin.mod.epinew)

summary(lin.mod.epinew)
plot(lin.mod.epinew)

ggplot(epi.results.sub, aes(x = population_log, y = EPI.new)) +
  geom_point() +
  stat_smooth(method = "lm")

ggplot(lin.mod.epinew, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs(title='Residual vs. Fitted Values Plot', x='Fitted Values', y='Residuals')
```

Try linear models with other variables!

Please push to your github repository:

1. All your code in a *.R or *.MD file
2. Boxplot comparing 3 variables
3. Q-Q plots for 3 variables compared to some known distribution
4. ECDF plots for 3 variables compared to each other
5. Summary stats and select plots from 3 linear models

Thanks!
Have a great weekend!