## Learning Objectives:

In this case study, together we will:

* Practice terminology and concepts related to correlation.
* Explore data fallacies and interpretation with respect to correlation and linear fitting.
* Practice R syntax and RStudio environment for data frames, 2D scatterplots, linear fitting, and correlation.

## Linear fit, slope, and correlationScatterplot showing a linear fit to data x-axis Number of stork breeding pairs y-axis Birth rate in thousands per year linear fit y = 0.0288x + 225.02 looks like a decent fit

Probably you have all seen some kind of scatter plot of data with a straight line drawn through it to indicate a relationship, in particular a **correlation**. That line is showing you a **linear fit of the data**. There are three key concepts that are associated with this: Fit, Linear, and r, which measures linear fit.

### Fit

**Fit** implies that there is a model that “fits” the data to the extent that the model predicts the data as much as possible. “As much as possible” is typically measured as the mean square error (MSE):



average of the squares of the difference between the data and what the model predicts

### LinearGraph with a variety of lines all going through (0,1), but with different slopes and their equations. Illustrates that positive slopes go up when reading left to right and negative slopes go down when reading left to right. It also shows how lines with slopes greater than 1 rise faster, and lines with positive slopes less than 1 rise slower. It also shows how lines with slopes les than -1 decline faster, and lines with negative slopes greater than -1 decline slower.

**Linear** refers to the type of model that is fit - a straight line of the form , where ***m*** is the **slope** of the line. See below for examples of how the slope impacts the steepness of that relationship. The error and/or “fit” have no impact on the equation of the model - a “best fit” line will be generated no matter how lousy it fits.

### Pearson’s Coefficient of Correlation, r

**Pearson’s Coefficient of Correlation, r,** is a measure of how much *x* and *y* are linearly related in a sample of data. Another interpretation for r is that **r2** (or simply **r-square**) is the amount of variance in the output that can be explained by the proposed model of the input.

* If **r2 = 1 (**or **r = -1 or 1),** the data are exactly in a straight line.
* If **r = 0,** a linear relationship does not explain the data at all.
* **Positive correlation (r > 0)** describes the situation when a larger *x*-value is related to a higher *y*-value.
* **Negative correlation (r < 0)** describes the opposite situation when a larger *x*-value is related to a lower *y*-value.
* You may notice that **r** is similar to the sign of the slope, ***m****,* of the linear model. That is because *m* has been multiplied by the ratio of the standard deviations to get **r**. However, the values are not the same.
* The r-value is also used to calculate **p-value**, which is a way to express the probability that this amount of correlation is generated by chance alone (p-values under 0.05 are generally considered “good enough” to provide evidence to reject the idea that there is no relationship).

### Discuss

This is the figure from the Bergstrom and West video.

This is a set of correlated points with various correlation values. In the first row we have points that cluster with correlation 1 (looks like a y=x line), 0.8 (same but grouped less closely), 0.4 (again same, but a lot less closely like a galaxy), 0 (almost a circular cloud of points), -0.4, -0.8, -1 ( all same as positive correlations but in the reverse direction so slopes down).
In the second row they show points very closely grouped as lines but with best fits that display different slopes, but with either 1 or -1 correlation values showing that while the + or - sign may be the same as the slope, the slope is NOT the same as correlation value. 
The last line is a bunch of 0 correlations with pretty pictures.

****Do the r-values here make sense to you?

****What is the difference between slope of the linear fit, ***m***, the Pearson correlation coefficient, **r,** and the **p-value**?

****What is the possible range of values the slope of the linear fit, ***m***, can take? The Pearson correlation coefficient, **r?** The **p-value**?

## Case Study Context - Do Storks Deliver Babies?

### MODELING/PROGRAMMING

**Modeling** is a broad term that encompasses a process including elements such as planning, implementation (“**mathematization” and/or coding**), testing, revision, feedback, etc., for some purpose just as programming does. The process of modeling has many starting points, for example from an idea, from an observation, from a set of data, or even from another model. In the case we are considering below, our process of modeling is:

1. Idea/Concept
 What question is being examined?
 Make a plan
2. Numerical
 Collect data
 Import and clean data
3. Visual
 Before analysis of data, see if a visualization of the data might give insights
4. Algebraic
 Fit a model to the data
 Use best fit model or established theory to choose a model
5. Verbal
 Interpret results of numerical, visual, and algebraic work


### REPRODUCING RESULTS

We know that storks and babies probably are not really related, but we will use this case study as a tool to investigate how correlation works and how interpretations can go wrong, with a dataset we know is correlated (thus reproducing the results of Matthews). Having reproducible results, we have learned in this class, is an important part of good science. In particular, programming in R gives us a code script that is extremely easy for someone else to transparently examine (the equations in spreadsheets are hidden), they can run the code on their computer, and in theory they should be able to “reproduce” the results of the study.

### PLAN

**\* Coding professional practice - Intentionally plan your code before you start typing.**

In order to determine whether there is a linear correlation between storks and babies, we will need to:

1. Store the data into variables (in this case the linear regression command operates on a data frame).
2. Plot the data in a scatterplot with axis labels & look to see if there is a reason to believe that a linear relationship exists (could be exponential, no relationship, etc.).
3. Code in R to perform a best fit analysis to determine the line of best fit & plot it.
4. View correlation and summary statistics of linear model.

### IMPLEMENT

**\*Coding professional practice - name the parameters close to what they mean. Use # (aka “comment out”) before descriptors that explain what you are doing and/or the units, or where the figure comes from.**

#### Step 0 - Prepare:

Start a new R-script (in RStudio, go to **File > New File > R-script**).

Add the file information to the top of your script.

#Diaz Eaton, DCS 105 A Calling Bull

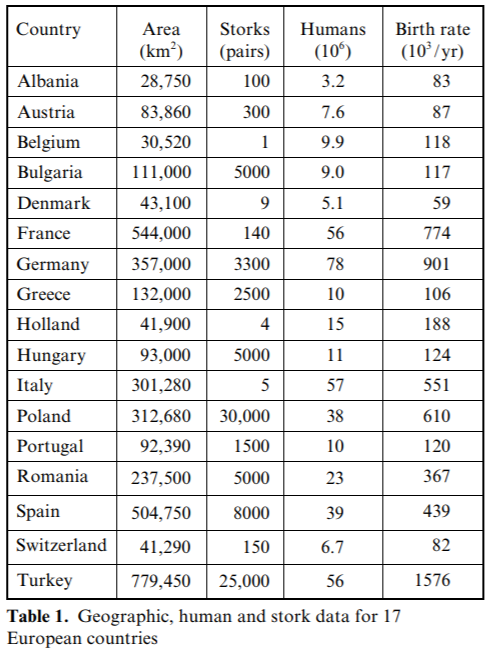
#Storks and Babies Case Study, Jan 29, 2019

#Based on Matthews

#### Step 1 (Copy and paste):

**\*Science Professional Practice - Usually in reproducing results, you can get an .xls file of the data to import, typically from the “Supplemental Material” section of a paper. In this case, because there isn’t much data, instead we will type it directly in (though you are lucky - you can just cut and paste the code below into your R script)!**

#Store the data into variables.

#country names

countries <- c("Albania", "Austria", "Belgium", "Bulgaria", "Denmark", "France", "Germany", "Greece", "Holland", "Hungary", "Italy", "Poland", "Portugal", "Romania", "Spain", "Switzerland", "Turkey")

#habitat in km^2

areakm <- c(28750, 83860, 30520, 111000, 43100, 544000, 357000, 132000, 41900, 93000, 301280, 312680, 92390, 237500, 504750, 41290, 779450)

#storks in pairs

storks <- c(100, 300, 1, 5000, 9, 140, 3300, 2500, 4, 5000, 5, 30000, 1500, 5000, 8000, 150, 25000)

#population data - number of humans in millions

humans <- c(3.2, 7.6, 9.9, 9.0, 5.1, 56, 78, 10, 15, 11, 57, 38, 10, 23, 39, 6.7, 56)

#number human babies born in thousands/year

births <- c(83, 87, 118, 117, 59, 774, 901, 106, 188, 124, 551, 610, 120, 367, 439, 82, 1576)

In this case the linear regression command operates on a data frame, so we have to arrange this information into a data frame.

#concatenate into data frame

Matthews\_df <- data.frame(areakm, storks, humans, births, row.names = countries)

**\*Coding professional practice - build in ways to check your code to make sure it is functioning as intended.**

Use str(Matthews\_df) in the command window/console to double check your data frame.

**** Does your data frame look right? How can you tell?

#### Step 2:

**#**Plot the data in a scatterplot with axis labels. Here is a [quick reference](https://www.statmethods.net/advgraphs/axes.html) of simple plot commands in R.

#scatterplot, also stored storks as x and births as y

plot(Matthews\_df$storks, Matthews\_df$births, xlab = "Storks (pairs)", ylab = "Birth rate (thousands/year)") # scatterplot

Look to see if there is a reason to believe that a linear relationship exists (could be exponential, no relationship, etc.). This is an important step - don’t skip this step (see ✨ below for a longer explanation of why).

**** Look at the scatterplot of data? Is there any reason to believe a linear correlation may exist between the two variables?

#### Step 3:

#Code in R to perform a best fit analysis to determine the line of best fit & plot it.

#linear regression

#syntax is lm([target variable] ~ [predictor variables], data = [data source])

lm\_sb = lm(births~storks, data = Matthews\_df)

abline(lm\_sb) #plots the line of best fit onto your data

 **Self-check -** Does the best fit line look like it fits the data?

#### Step 4:

#View correlation, summary statistics, and the equation of linear model.

#Examine the output

summary(lm\_sb) #output

cor(births, storks) #try switching storks and births for fun

 **Self-check -** What result did you get? Same as Matthews? Different? Is the code working properly?

#Code that extracts the fit coefficients and displays the equation on your plot

lm\_coef <- round(coef(lm\_sb), 3) # extract coefficients

**** How is this line of code extracting coefficients?

**\*Coding Professional Practice Hint:** Try using the console to run coef(lm\_sm) to see what happens. Then type help(round) to see what the round function does and how it works. Breaking up the command line into smaller pieces helps us better understand what the code is doing.

# display equation

mtext(bquote(y == .(lm\_coef[2])\*x + .(lm\_coef[1])),

adj=1, padj=0)

**** How is this line of code displaying the equation for the line?

### ANALYZE

****What is the:

Correlation coefficient

slope of the best fit line

r-squared (note there are 2 values)

p-value

equation of the best fit linear model?

**Do storks deliver babies? Why or why not?**

### MODIFY

1. **Discuss & Plan:** The alternative explanation that Bergstrom & West provide is that area explains both. How would you modify the code you wrote in the script to check if a relationship between two different variables of the matrix might exist (if area is a confounding variable)?
2. **Implement** your modification

**\*WARNING - Professional practice**: start a new .r file and use new storage variables so you don’t write over old, working code and make sure you implement the self-checks above.

1. **Analyze output:** 
   1. What answer did you get?
   2. What does it mean?

**Do storks deliver babies? Why or why not?**

**SHOW ME (or the TA) YOUR NEW PLOTS AND R2 VALUES FOR COMPLETION! And also be prepared to answer a question about the final results or one of the s.**

### CHALLENGE (Optional)

✨ I mentioned that it is important to check the scatterplot before just running a linear fit. We also typically examine a plot of the residuals as well. Here’s Anscombe’s quartet - the pinnacle example of why:

<https://en.wikipedia.org/wiki/Anscombe%27s_quartet>

Explore this idea on your own by following this tutorial: Stats blog by Sean Dolinar <https://stats.seandolinar.com/introduction-to-correlation-with-r-anscombes-quartet/>)

And explore the Datasaurus Dozen for even cooler examples

<https://www.autodeskresearch.com/publications/samestats>

## References

Do Storks Deliver Babies? By Matthews

<http://robertmatthews.org/wp-content/uploads/2016/03/RM-storks-paper.pdf>

Reference of simple plot commands in R

<https://www.statmethods.net/advgraphs/axes.html>

For a more full introduction to regression with R, check out this blog, which has links to relevant DataCamp tutorials.

<https://www.datacamp.com/community/tutorials/linear-regression-R#what>