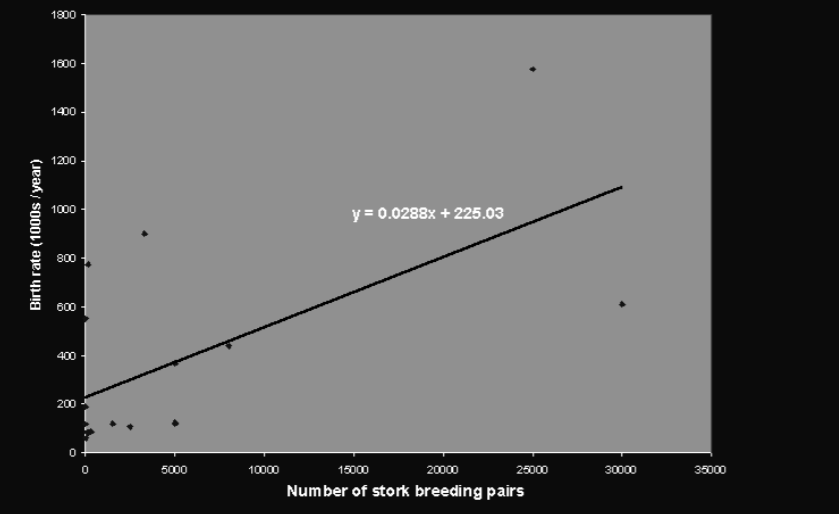
**Storks Deliver Babies Case Study**

An introduction to correlation fallacies with R, inspired by [Calling Bull](https://callingbullshit.org/) and Matthews “[Storks Deliver Babies](http://robertmatthews.org/wp-content/uploads/2016/03/RM-storks-paper.pdf)” by Dr. Diaz Eaton, Bates College, for use in DCS 105A.

Learning Objectives:

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



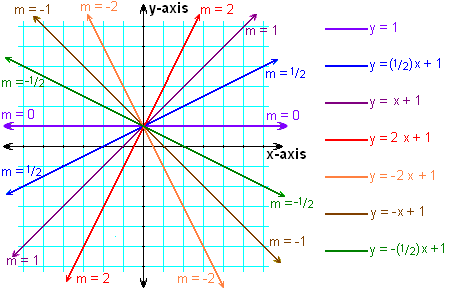
**Linear fit, slope, and correlation**

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** 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

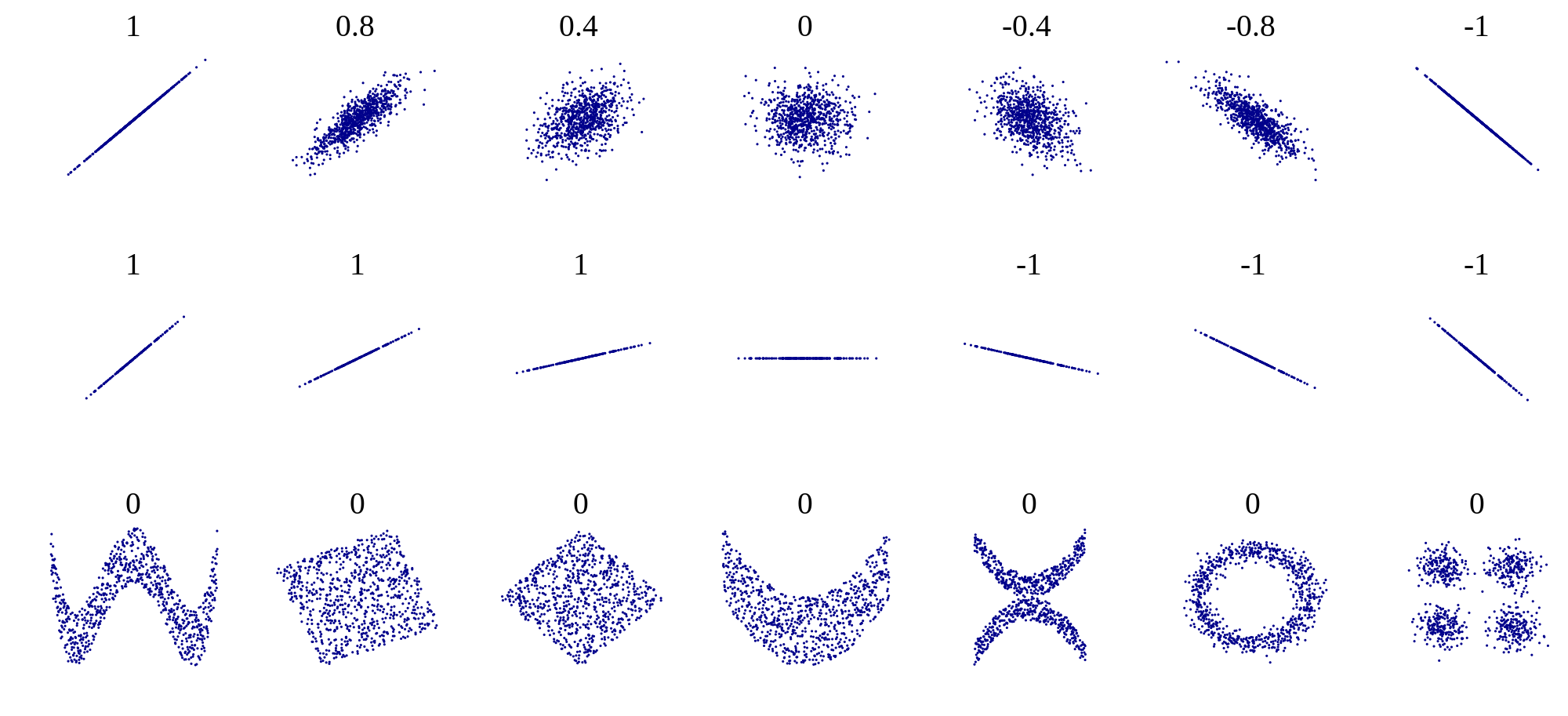


* **Linear** implies the type of model that is fit - a straight line of the form , where *m* is the **slope** of the line. See below for examples of how the slope impacts the steepness of that relationship. The error and/or “fit” have no impact on the equation of the model - a “best fit” line will be generated no matter how lousy it fits.
* **Pearson’s Coefficient of Correlation, r,** 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 this 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 video. Do the r-values here make sense to you?



**MODELING/PROGRAMMING**

**Modeling** 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. Occasionally the predicted linear model is not just a mathematical description of the data points, but is used to make some inference about the nature of the relationship. 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:

abstract code code (with underlying math) interpret

Idea/concept → numerical data → visual plot → symbolic representation → verbal

**Storks Deliver Babies**

We know that storks and babies probably are not really related, but we will use it just to get an idea about how correlation works, with a dataset we know is correlated (thus reproducing the results of Matthews - [citation](http://robertmatthews.org/wp-content/uploads/2016/03/RM-storks-paper.pdf)).

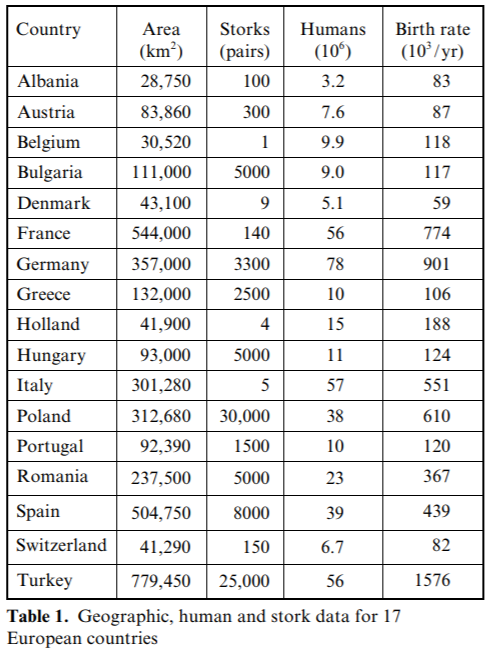
**PLAN**

**\* 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.** 

**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:**

#Store the data into variables (in this case the linear regression command operates on a data frame).

#habitat and storks

areakm <- c(28750, 83860, 30520, 111000, 43100, 544000, 357000, 132000, 41900, 93000, 301280, 312680, 92390, 237500, 504750, 41290, 779450) #km^2

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

#millions of babies born - population data

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) #in millions

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

#concatenate into matrix

Matthews\_df <- data.frame(areakm, storks, humans, births)

#Make row and column names

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

row.names(Matthews\_df) <- 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 to double check your data frame.

**Step 2:**

**#**Plot the data in a scatterplot with axis labels. Here is a [quick reference](https://www.statmethods.net/advgraphs/axes.html) of 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 an longer explanation of why).

Discuss your plot.

**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

**Step 4:**

#View correlation and summary statistics of linear model.

#Examine the output

summary(lm\_sb) #output

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

**ANALYZE**

What result did you get? Same as Matthews? Different? Is the code working properly?

What is the:

Correlation coefficient

slope of the best fit line

r-squared (note there are 2 values)

p-value

What is the equation of the line?

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

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

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

adj=1, padj=0) # display equation

**MODIFY**

1. **Discuss & Plan:** The alternative explanation that J&W 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?
2. **Implement**
3. **Analyze output:** 
   1. What answer did you get?
   2. What does it mean?

**CHALLENGE (Optional)**

✨I mentioned that 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>

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>