## Correlation and Regression

This course examines bivariate relationships.

* Both variables are numerical
* The y or dependent variable is referred to as the response variable
* The x or independent or predictor is something you think might be related to the response
* A scatterplot is the best way to visualize a bivariate relationship
  + form (e.g. linear, quadratic, non-linear)
  + direction (positive or negative)
  + strength (how much scatter / noise) as quantified by magnitude of the correlation
  + outliers
* Sometimes carefully transforming one or both variables can reveal a clear relationship
* A boxplot is basically a scatterplot in which the independent variable has been discretized

Basic scatterplot syntax 🡺 ggplot(ncbirths, aes(y = weight, x = weeks)) + geom\_point()

ggplot(noise, aes(x, y)) + geom\_point() + facet\_wrap(~z)

Basic boxplot syntax 🡺 ggplot(ncbirths, aes(y = weight, x = cut(weeks, break = 5))) +  
 geom\_point()

Basic transformation syntax 🡺 ggplot(mammals, aes(y = BrainWt, x = BodyWt)) +  
 geom\_point() +

coord\_trans(x = “log10”, y = “log10”)

two different approaches

scale\_x\_log10() +

scale\_y\_log10()

**Correlation and Correlation Coefficient (Pearson product-moment correlation)**

* The direction of the relationship is indicated by the sign
* The strength of the relationship is quantified by the magnitude
* Used to assess the strength and direction of a **linear bivariate** relationship
* Correlation does not imply causation
* Spurious correlation are remarkable but nonsensical movements in two variables; **time** is oftentimes a confounder; when you see two variables compared across time, beware of the role of time being a potential confounder; **space** can also have a confounding effect

Basic correlation coefficient syntax 🡺 ncbirths %>%

summarise(N = n(), r = cor(weight, mage))

add use = “…”

use = “pairwise.complete.obs”

**Simple Linear Regression**

dfg