## Correlation and Regression

Correlation and Regression deals with ***bivariate relationship*** in which both variables are numerical.

* The y or dependent variable is referred to as the response.
* The x or independent variable or predictor is something you think might be related to the response
* A scatterplot is one of the best ways to visualize bivariate relationships with respect to:
  + form (linear, quadratic, non-linear, etc…)
  + direction (positive or negative)
  + strength (how much scatter / noise)
  + outliers
* Sometimes carefully transforming one or both variables can reveal a clear relationship
* A boxplot is basically a scatterplot in which the predictor has been discretized

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

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

Basic boxplot 🡺 ggplot(ncbirths, aes(y = weight, x = cut(weeks, breaks = 5))) +  
 geom\_boxplot()

Basic transformation 🡺 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 of the correlation coefficient
* The strength of the relationship is quantified by the magnitude of the correlation coefficient
* The correlation coefficient is used to assess **linear bivariate** relationships
* Correlation does not imply causation
* Spurious correlation are remarkable but nonsensical movements in two variables; **time** is often a confounding variable; when you see two variables compared across time, beware of the potential confounding role of time; **space** can also have a confounding effect

Correlation coefficient 🡺 ncbirths %>% summarise(N = n(), r = cor(weight, mage))

add use = “…”

use = “pairwise.complete.obs”

**Simple Linear Regression (SLR)**

SLR is a specific example of a larger class of smoothing models. SLR finds the “best fit” line that cuts through the data in a way that minimizes the distance between the line and the data points. Properties of the least square algorithm that SLR uses to find the best fit line includes:

* Easy, deterministic, and unique solution
* Residuals guaranteed to sum to zero
* Best fit line must pass through



response = *f*(explanatory variable) + noise

Basic SLR 🡺 ggplot(data = bdmins, aes(y = wgt, x = hgt) +

geom\_point() +

geom\_abline(slope = 1, intercept = 0) +

geom\_smooth(method = “lm”, se = FALSE)

Linear Model Object 🡺 lm(Performance ~ SP500, data = pfg)

Various functions can be used to extract information from an lm object:

* coef()
* summary()
* fitted.values() Note: mean(response) = mean(fitted values) in SLR
* residuals() Note: mean(residuals) = 0 in SLR