**Correlation & Regression**

* **Correlation** – statistical method used to determine whether a relationship exists between variables
* Consider r. v. X, Y
  + X – independent variable; Y – dependent variable (response)
* Correlation coefficient – measures the strength & direction of a linear relationship between 2 variables
  + Population correlation = ρ
    - ρ = SXY/√(SXX ⋅ SYY )
    - -1 ≤ ρ ≤ 1
    - ρ → 1 = strong positive correlation
    - ρ → -1 = strong negative correlation
    - ρ → 0 = weak correlation
    - If X, Y are independent, then ρ = 0 (no correlation)
    - If ρ = 0, X, Y are uncorrelated (not necessarily independent)
  + Sample correlation = r
* Hypothesis testing for the correlation coefficient
  + Step 1 – formulate null & alternate hypotheses
    - H0: ρ = 0
    - Ha: ρ ≠ 0
  + Step 2 – compute test value
    - ; d.f. = n – 2
  + Step 3 – find p-value/critical value
  + Step 4 – reject/do not reject H0
* **Regression** – statistical method used to describe the nature of the relationship between variables
* Regression line/line of best fit – the sum of the squares of the vertical distances between each point & the line is minimum
  + i.e. method of least squares
  + Least squares estimates:
    - a = y-bar – b ⋅ x-bar
    - b = Sxy/Sxx
  + Fitted regressions line:
    - y^ = a + bx
  + Residual = observed value – fitted value
    - εi = yi – a – bxi
  + Residual sum of squares
    - SSE = ∑ ε = Syy – Sxy2/Sxx
  + The above estimates for a, b minimize the SSE