Big Data for Cities

Week 9

Curt Savoie
Connor McKay

Agenda

- Recap of last week
- Advanced Statistics / Regression
- R Demo

Recap

- Location based analysis?
- R issues?

More Statistics!

- Correlation
- Regressions
- ANOVA

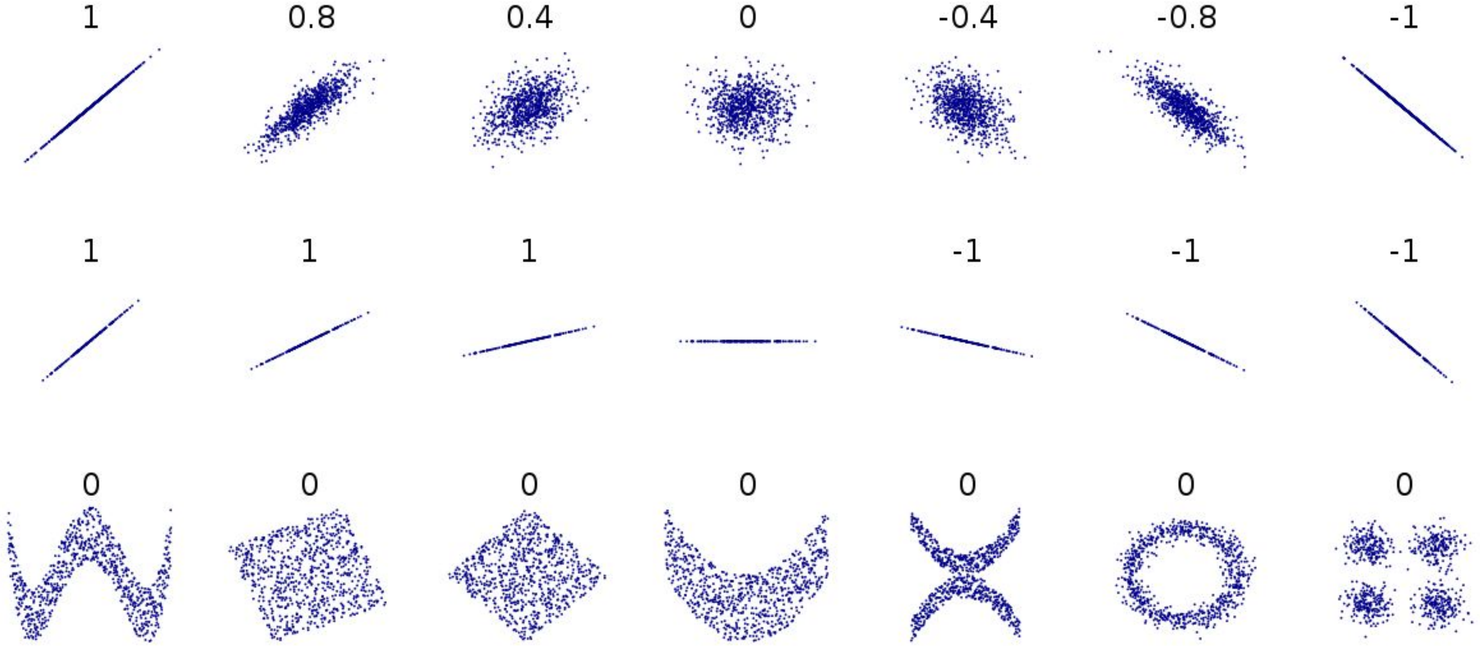
Linear Regression!

- **What is it?**
  - is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable.
  - Slope = 0, no linear relationship
- **When do you use it?**
  - The relationship between the variables is linear.
  - Normal Distribution

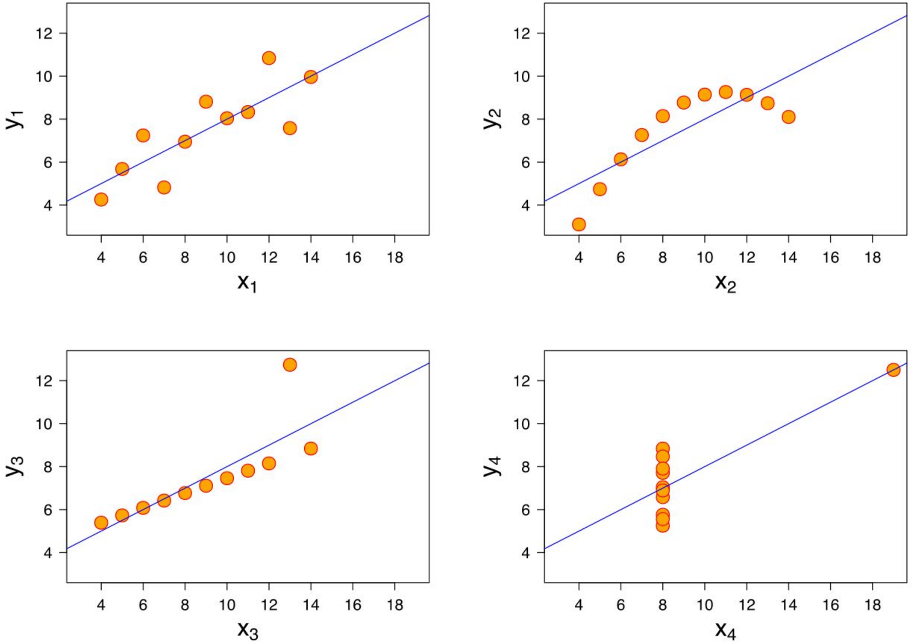
Linear Regression!

- Caveats
    - can be affected by data clustering
    - does not accurately describe nonlinear relationships
    - can be affected by ’outlier data points’,

Correlation!



Correlation! (Not all plots are equal!)



ANOVA!

- What is it?
    - Analysis of Variance
    - compares the means of two or more independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different
    - F test

ANOVA!

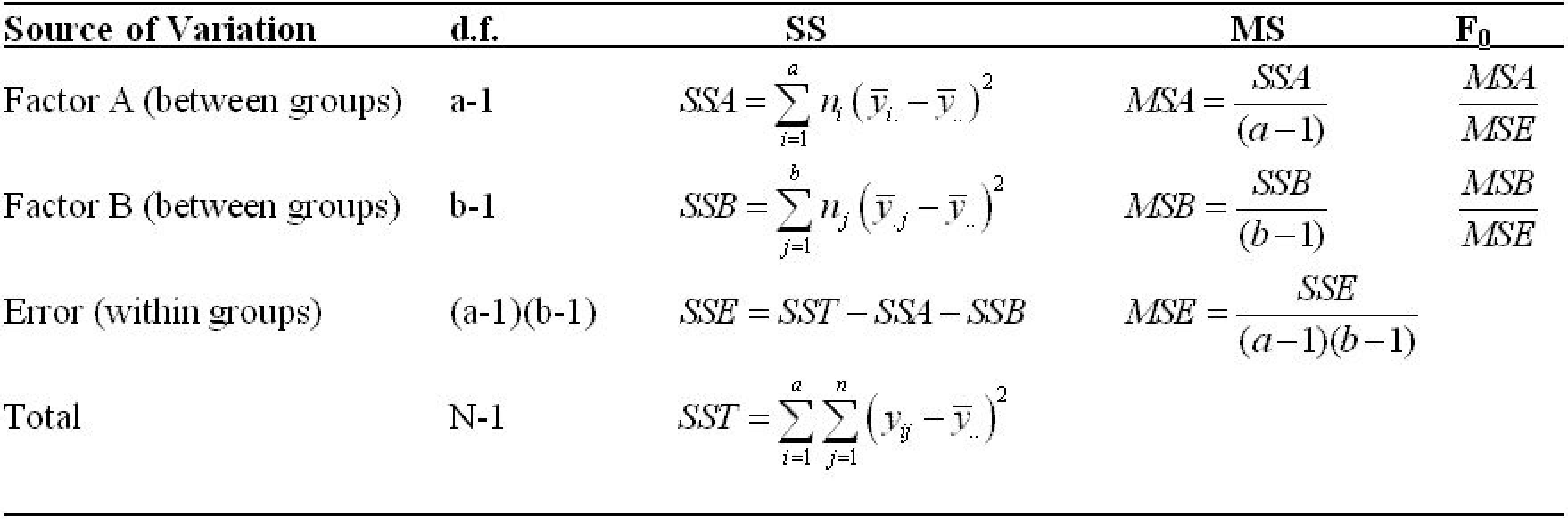
- **When do you use it?**
    - Statistical differences among the means of two or more groups
    - Statistical differences among the means of two or more interventions
    - Statistical differences among the means of two or more change scores

ANOVA! (data requirements)

- Dependent variable that is continuous (i.e., interval or ratio)
- Independent variable that is categorical (i.e., two or more groups)
- Cases that have values on both the dependent and independent variables
- Independent samples/groups (i.e., independence of observations)
    - There is no relationship between the subjects in each sample. This means that:
        - subjects in the first group cannot also be in the second group
        - no subject in either group can influence subjects in the other group
        - no group can influence the other group
- Random sample of data from the population
- Normal distribution (approximately) of the dependent variable for each group

ANOVA!

SS: Sum of Squares, d.f.: degrees of freedom MS: Mean Square



Source of Variation	d.f.	SS	MS	F_0
Factor A (between groups)	a-1	$SSA=\sum_{i=1}^{a} n_i\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2$	$MSA=\frac{SSA}{(a-1)}$	$\frac{MSA}{MSE}$
Factor B (between groups)	b-1	$SSB=\sum_{j=1}^{b} n_j\left(\bar{y}_{.j}-\bar{y}_{..}\right)^2$	$MSB=\frac{SSB}{(b-1)}$	$\frac{MSB}{MSE}$
Error (within groups)	(a-1)(b-1)	$SSE=SST-SSA-SSB$	$MSE=\frac{SSE}{(a-1)(b-1)}$	
Total	N-1	$SST=\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{..}\right)^2$		

For Next Week

- **Reading on theory and practice**
    - https://projects.fivethirtyeight.com/p-hacking/
    - FOR 2 WEEKS FROM NOW
        - https://en.wikipedia.org/wiki/Association_rule_learning
        - https://en.wikipedia.org/wiki/Network_theory
- In R
    - More homework