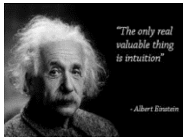
Transformation #10:
Data Reduction
(Feature Extraction)

XI(✖) – X’s are not independent (are correlated)



Data Reduction: Intuition

- Too many predictors in a model usually suffer from **over-identification** and **multicollinearity**, particularly if some variables are correlated
- But some times business knowledge suggests that all or many of these predictors **do belong** in the **model**
- Business models generally don't include too many variables, but models in **biology** and other fields can have **thousands** of variables
- One way to resolve this issue is to develop a **"structural model"** (multiple models estimated together -- will cover later in the semester)
- Another way is with **data reduction** or **feature extraction** methods
- This involves **combining** groups of (usually correlated) variables into **factors** or **latent variables**, either through **aggregation** or linear combination of variables into larger variables
- Popular data reduction methods include: **factor analysis** (FA – often used with **survey** data), **principal components analysis** (PCA) and **partial least squares** (PLS)
- We will cover these methods later in the semester

R Tips

See lecture on Principal Components Regression (PCR) and Partial Least Squares (PLS) Regression

Other Transformations

- There are **endless options** for data transformations in pre-processing. We have covered the most popular ones.
- Examples of other transformations:
  - **Re-scaling:** e.g., from ^{0}F to ^{0}C, mpg to kpg
  - **Reverse scaling:** often used to facilitate interpretation – e.g., a 1-7 satisfaction scale can be converted into a dissatisfaction scale by subtracting the value from 8, so that a 1 becomes 7 and a 7 becomes 1
  - **Inverse:** $\boldsymbol{x^* = \frac{1}{x}}$ similar purpose than revers scaling, but this is a non-linear transformation, harder to interpret, and x cannot be 0
  - **Logit:** we will discuss this in depth later for classification models, but the Logistical regression is simply a transformation of the dependent variable using the logistic function.

Kogod School
of
Business