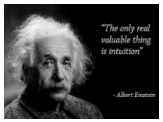
Survey Data



Survey Data: Intuition

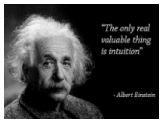
- Surveys are a most **popular** way to collect data
- The variety of survey **designs** and **methods** is quite broad (see: http://www.ats.ucla.edu/stat/mult_pkg/faq/svy_howtochoose.htm for popular designs, methods and **R libraries**)
- For this class, we focus on survey data with **simple random design**, which are widely used to collect marketing, opinion and other trend data, with a large number of questions, aimed at predicting outcomes.
- It is not uncommon to collect **100+** survey **items**, which presents several **problems** for data **analysis**
- The most frequent types of **analysis** include **simple statistics** like means, variance and frequencies
- But as we will see in this section, there are very **sophisticated methods** to extract interesting **meaning** out of survey data

Survey Data Collection Issues

- Survey data analysis presents **unique problems** at both stages: data collection and analysis
- This is **not** a survey **methods** class, so we will not cover data collection issues in detail, but it is important to collect data that will **meet** the **assumptions** of the analysis models to be used, so let's look at three **OLS** assumptions that matter for survey models:
  - ✓ (**YC**) **Y is continuous** – It is important to define the survey outcome variables of interest, which will determine the type of predictive modeling to use: Continuous (e.g., sales) → OLS, GLM; Count (e.g., number of customers) → GLM, Poisson; Binary (e.g., in-favor vs. against) → Logistic, LDA, Trees
  - ✓ (**XI**) **X's are independent** – need to think of dimensionality and collinearity when designing survey instruments
  - ✓ (**OI**) **Observations are independent** – random sampling is key → e.g., if you ask the same question to two close friends or relatives you are likely to get biased responses

Survey Data Analysis Issues

- A simple random design survey with a large number of survey items collected can present many **issues** for data **analysis**. Some of the most commonly found problems are:
  - ✓ **Dimensionality** – Because the data set has many items, it is important to select analysis methods that address dimensionality problems, particularly collinearity → structural equations, shrinkage models, PCA, PLS, etc.
  - ✓ **Common Method Variance** – When respondents answer both outcome and predictor questions the answers are likely to be correlated (e.g., Do you like your job? Are you happy? → Job satisfaction predicts happiness)
  - ✓ **Concurrent/Discriminant Validity** – Ensuring that the survey items measure what we intend to measure (concurrent) and that the outcome is differentiated from the predictors (discriminant)
  - ✓ **Reliability** – we can group correlated survey items into more general predictors, as long as they are correlated as a group – i.e., reliability is high



Dimensionality Issues

- What do you do if you have several survey items that are highly correlated, exhibiting severe multicollinearity, but you think they are important predictors that need to be modeled? Popular approaches:
  - Subset Selection – i.e. find a reduced set of survey items to avoid multicollinearity → throwing away survey data
  - Structural Equations – i.e., a group of (i.e., more than one) interrelated regression models → Lisrel, PLS, Hierarchical OLS
  - High dimensional methods – e.g., shrinkage, PCA, PLS → not very useful for interpretation, which is often the goal
  - Factor Analysis – i.e., a data reduction method uniquely suited for survey (and other type of high-dimensional) data analysis → PCA, Varimax Rotation

KOGOD SCHOOL
of
BUSINESS