Analysis of Variance
Part 1: Principles of Experimental Design

STAT 705: Regression and Analysis of Variance

Introduction

Two Possible Sources of Data		
	Observational Study	Designed Experiment
Randomization?		
Predefined treatments?		
Scope of inference?	Association	Causation

Both types of data can be analyzed with linear models

Observational Studies

- Data are recorded without interfering with the course of events, i.e., the researcher is a passive observer
- Example: Traffic vs. Lead concentration (from our very first lesson)
  - Locations along the highway are randomly selected
  - Traffic volume (predictor variable) is observed
  - Lead concentration in tree bark (response variable) is measured
  - There is no attempt to control or manipulate these variables
- Other examples: Opinion polls, ecological studies
- Scope of inference
  - Can determine association, but not causation

Experiments

- Researchers actively intervene to control the study conditions and record the responses
- Values of predictor variables (called ‘factors’) are systematically altered for the purpose of measuring the resultant change in the response
  - Pre-defined values for the factors are called ‘levels’
  - A ‘treatment’ is a specific combination of the levels of the factors
- Experimental Units (EUs) are the entities to which the treatments are applied
- Can infer causation (i.e. determine if changes in response are CAUSED by changes in factors)

Common Designs

- One-way design
  - One factor, with k levels
  - Each level is a treatment
- Multi-way factorial design
  - Multiple factors
  - The treatments consist of every possible combination of the levels of the factors
- Block Designs
  - A block is factor created by the researcher to help explain variation in the response

Experimental Error

- Treatments should help explain some of the total variability in the response
- Excess variability (not accounted for by the treatments) is experimental error
- Possible sources of experimental error
  - Natural variation among EUs
  - Variability in measurement of response
  - Failure to include factors that affect the response
  - Inability to exactly reproduce experimental conditions from one EU to another
- Reducing the experimental error is VITAL for inference

Randomization

- A key component of any designed experiment
- Treatments and EUs are linked by random assignment
  - Which EU gets which treatment is decided randomly
  - This provides reasonable assurance that any difference we may see in the response variable is due to the treatments and not due to differences between EUs
- In practice, this can happen two ways
  - When the researcher creates the conditions that define the treatments, then the EUs are randomly assigned to a treatment
  - When the conditions that define the treatments are naturally occurring characteristics (e.g. gender, species), then the EUs are randomly selected from the population that has those characteristics

How to Perform Randomization

- For small studies
  - Create pieces of paper, one for each EU
  - Put the paper in a container and shuffle them
  - Without looking, repeatedly select a piece of paper
    - » EU on the 1st paper is assigned to the 1st treatment
    - » EU on the 2nd paper is assigned to the 2nd treatment
    - » Cycle through the treatments until all EUs are assigned
- For larger studies
  - Use computer-generated random numbers instead of pieces of paper

Replication

- Replication is the number of independent repetitions of a treatment
  - i.e., the number of EUs assigned to each treatment
- Without replication
  - We cannot estimate the experimental error
  - We cannot estimate the variability of the treatment effects
  - We have no way of measuring how ‘typical’ our results are

Balanced Designs

- Balanced designs
  - Have the same number of EUs in every treatment
  - Provide for more precise inference
- Typically, data sets are not balanced. Missing data can occur for a variety of reasons
  - Laboratory equipment failure
  - Omissions in recording the data
  - Personnel issues, weather-related issues, etc.
- Start with a *design* that is balanced, and try to avoid missing data

Example: Motor Oil Containers

Plastic containers for motor oil are blow molded in a machine that has two feeders, each feeding into three molding stations (for a total of six stations). Plastic is extruded through the feeders into continuous cylinders between two halves of each mold. As the molds close, the cylinders are pinched off and air is blown into them to form the container shapes. The production engineer is concerned that the molded containers at different stations are not the same weight. To investigate whether or not this is true, a random sample of 8 containers is taken from each station.

(Source: Tamhane and Dunlop, 2000)

Components of Oil Example

- This is a one-way design
  - A single factor (Station), with 6 levels
  - Six treatments (the six stations)
- Experimental Unit: a container
- Replications
  - 8 containers per station
  - 6 stations and 8 containers per station = 48 total observations
- Response variable: weight of container (in grams)

Example: Cloth Dyeing Experiment

The quality control department of a fabric finishing plant is studying the effect of several factors on dyeing for a cotton cloth used to manufacture shirts. Two operators (1 or 2), three cycle times (40, 50 or 60 minutes) and two temperatures (300° or 350°) were selected, and three small specimens of cloth were dyed under each set of conditions. The finished cloth was compared to a standard, and a numerical score was assigned.

Components of Dyeing Example

- This is a 3-way factorial experiment
- Three factors
  - Operator, 2 levels
  - Cycle time, 3 levels
  - Temperature, 2 levels
- Treatments
  - total 2×3×2 = 12 treatments
- Replications
  - 3 replications per treatment
  - 3 × 12 = 36 total observations
- Experimental Unit (EU)
  - small specimen of cloth
- Response variable : numeric score

Treatment	Operator	Cycle Time	Temp.
1	1	40	300
2	1	40	350
3	1	50	300
4	1	50	350
5	1	60	300
6	1	60	350
7	2	40	300
8	2	40	350
9	2	50	300
10	2	50	350
11	2	60	300
12	2	60	350

What You Should Know

- Recognize the difference between observational studies and experiments
- For experiments, be able to identify
  - factors
  - treatments
  - experimental units
  - number of replications
  - number of observations
  - response variable