Stat 428/528: Advanced Data Analysis 2

Chapters 4: One Factor Designs and Extensions

February 6, 2019

ANOVA

- We think of regression as having a quantitative response and (usually) quantitative predictors
- ANOVA has a quantitative response and *qualitative* predictors. ——ANOVA is a really a special case of regression.
- The regression framework can handle qualitative predictors and mixes of qualitative and quantitative predictors.
- We'll spend some time in the ANOVA setting, where all predictors are qualitative, before going back to the general regression setting.

ANOVA

Often ANOVA arises in designed experiments

- the experimenter decides certain conditions to be manipulated. A lot of concepts from ANOVA historically came from agricultural experiments
  ——where different growing conditions were randomly assigned to different plots of land to see which farming techniques affected the yield of the crop.
  —— Variables that could be manipulated might include things like watering regimes and type of fertilizer used.
- In a greenhouse, experimenters can also control things like temperature and humidity. Whether or not these are considered qualitative or quantitative can depend on the design of the experiment. In many cases, experiments designed as ANOVAs just use high and low values for variables that could be treated as quantitative, such as temperature.

ANOVA

The phrase **completely randomized one-factor design** is used to refer to an experiment in which only one primary factor (i.e., treatment) is analyzed, and this factor or treatment is randomly assigned to each individual.

The experimenter can decide how many times each treatment is given.

- if an experiment has 3 levels (treatment A, treatment B, and placebo) for blood pressure
- there are 30 subjects
- the experimenter can randomly choose 10 patients to receive treatment A, then randomly pick 10 patients from the remaining 20 to receive treatment B, then give the placebo to the remaining 10 subjects.

ANOVA

This randomization is more like distributing cards from a shuffled deck (where you sample without replacement) than rolling a die (sampling with replacement). This randomization could be accomplished in R as follows, assuming that patients are numbered with IDs 1 through 30:

```
> patient <- 1:30
> treatment <- c(rep("a",10),rep("b",10),rep("placebo",10))
> treatment <- sample(treatment,replace=F)
> mydata <- data.frame(cbind(patient,treatment))
> head(mydata)
  patient treatment
1       1         b
2       2   placebo
3       3   placebo
4       4         a
5       5   placebo
6       6         a
```

ANOVA

Balanced ANOVA or balanced design

- $n = 30$, $n_g = 10$ for each group with different treatment, A, B and placebo
  ——equal sample sizes in each group.
- This is usually preferable in terms of statistical power.
  —— In other words, if there are different means for the groups, then you have a higher probability of detecting those differences using equal sample sizes than using unequal sample sizes if other assumptions are met (such as equal variances).

ANOVA

Comments:

- Differences in blood pressure could be due to a number of unmeasured variables such as age, sex/gender, initial blood pressure, genetic influences to responses to the drugs, differences in lifestyle etc.
- By doing random assignment, the different groups will be similar in terms of the distribution of age, sex, genetics, etc.
- Alternatively, the experimenters could try to make the population sampled from more uniform by only recruiting people from one sex, one age range, etc.
- In summary, an experiment is to impose a treatment on experimental units to observe a response. Randomization and carefully controlling factors are important considerations.

One way ANOVA model

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

- y_{ij} refers to the jth individual in the ith treatment group.
  —— For the blood pressure example, we would have $j = 1, \ldots, 10$ and $i = 1, 23$.
  —— The quantity $y_{2,5}$ for example, would mean the 5th individual receiving treatment B
- μ_1, μ_2 and μ_3 are the population mean for all potential responses to the ith treatment.
  —— An individual in group 2 has an expected blood pressure of μ_2
- $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$
  ——The responses within and across treatments are assumed to be independent, normal random variables with constant variance.
  —— $\varepsilon_{2,j}$ represents the deviation of the jth individual in group 2 from μ_2.

ANOVA

Define μ as the **Grand mean** and α_i as the **treatment effect**.

$$\mu = \frac{1}{I}\sum_{i=1}^{I}\mu_i, \alpha_i = \mu - \mu_i$$

The treatment effects measure the difference between the treatment population means and the grand mean and are constrained to add to zero,

$$\alpha_1 + \alpha_2 + \cdots + \alpha_I = 0.$$

Another way to write the ANOVA model is

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

Response = Grand Mean + Treatment Effect + Residual.

ANOVA

The null hypothesis for ANOVA is that

$$\mu_1 = \mu_2 = \cdots = \mu_I$$

where I is the number of groups. Or

$$\alpha_1 = \alpha_2 = \cdots = \alpha_I$$