Theory on Covariate Adjustment

March 16, 2022

0.1 Introduction

This section is based on "The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition"

The main advantages of including the covariate in randomized experiments are:

1. generally greater power (in RCT is this the major payoff)
2. a reduction in bias caused by chance differences between groups that exist before the treatments are administered (this bias is generally small in randomized designs)
3. conditionally unbiased estimates of treatment effects

Types of covariates:

- baseline measures
- organismic characteristics
- environmental characteristics

Conditional versus Unconditional Inference

In expectation, random assignement provides groups that are exactly equivalent. So it follows that the difference between sample means on Y (after treatment) is an unbiased estimate of the treatment effect because the groups are in expecation the same the only difference between the groups is due to the treatment.

But in the case of a single experiment the two groups will never be exactly the same on a continuous variable before treatments are applied. This means if we use the difference between sample means on Y gives us a conditionally biased estimate of the treatment effect because groups were not exactly the same before treatment. This does not mean that an ANOVA F-test and the associated effect estimate are wrong, but it is possible to do better.

ANCOVA incorporates the information available on the X variable and provides a conditionally unbiased (conditional on X) estimate of the treatment effect.

The long run average of these two types of effect estimates are the same under random assignement, but the ANCOVA estimates are more precise.

General Ideas Associated with ANCOVA

ANCOVA will statistically partition the effect of the covariate measure from the relationship between the treatments and the dependent variable.

The ANCOVA F-test is more likely to identify a statistically significant treatment effect than the ANOVA F-test.

- ANOVA error term is based on variation of Y around individual group means.

$$\hat{\epsilon}_{i;group\,j} = (Y_{ij} - \bar{Y}_j) \tag{1}$$

- ANCOVA error term is based on variation of Y scores around pooled within group regression lines

$$\hat{\epsilon}_i = (Y_{ij} - \hat{Y}_{ij}) \tag{2}$$

ANOVA and ANOVAR

ANOVA is generally used to test the equality of population means, ANOVAR is used to test the hypothesis of zero population slope. One can look at ANCOVA as an integration of ANOVA and ANOVAR, but one can also look at it as a minor variant of multiple regression analysis.

0.2 Analysis of Covariance Model

The statistical model for the analysis of covariance is:

$$Y_{ij} = \mu + \alpha_j + \beta_1(X_{ij} - \bar{X}..) + \epsilon_{ij} \tag{3}$$

where

Y_{ij}	is the dependent variable score of the ith individual in the jth group
μ	is the overall population mean (on dependent variable)
α_j	is the effect of treatment j
β_1	is the lienar regression coefficient of Y on X
X_{ij}	is the covariate score for the ith individual in jth group
$\bar{X}..$	is the grand covariate mean
ϵ_{ij}	is the error component associated with the ith individual in the jth group

0.3 Computation and Rationale

We write here:

$$N = total number of subjects$$
$$n_1, ..., n_k the number of subjects in groups 1, .., k$$
$$X = the value of the covariate$$
$$x = X - \bar{X}..$$
$$Y = value of the dependent variable$$
$$y = Y - \bar{Y}$$

1. Computation of total sum of squares: They consist of the treatment effects, differences that can be predicted from the covariate X and differences that are neiter (i.e., error)

$$TotalSS = \sum y_t^2 = \sum Y_t^2 - \frac{1}{N}(\sum Y_t)^2 \tag{4}$$

2. Computation of total residual SS

$$SS(SSres_t) = \sum y_t^2 - \frac{(\sum xy_t)^2}{\sum x_t^2} = \tag{5}$$

$$[\sum Y_t^2 - \frac{1}{N}(\sum Y_t)^2] - \tag{6}$$

$$[\frac{[\sum XY_t - \frac{(\sum X_t)(\sum Y_t)}{N}]^2}{\sum X_t^t - \frac{(\sum X_t)^2}{N}}] \tag{7}$$