The Logic of Controls and Comparisons in Social Science

Nicola Righetti

Summary

- In this lesson, we analyze the role of comparative logic in social research.
- We will see this logic implemented in several research designs.
- In particular we will see the experiment, comparative research (Most Similar and Most Different System Designs) quasi-experiments including natural experiments (Pretest-Posttest Quasi-Equivalent Groups design, Interrupted Time Series design, and Difference in Difference design)

The Logic of Controls and Comparisons and Social Research

- In social research, the logic of controls and comparisons is central to establishing relationships between variables and understanding social phenomena.
- This logic involves methodically comparing different groups or conditions to isolate the effect of a particular variable.
- The effort to isolate the effect of variables, utilizing various techniques, is applicable to both qualitative (e.g., comparative qualitative studies) and quantitative research designs (e.g., using regression analysis, statistical comparisons, experiments).

Comparative vs Non-Comparative Research

The methodology of social science is inherently comparative.

- The term *comparative method* typically is used in a narrow sense to refer to the branch of social science concerned with cross-societal differences and similarities.
- However, continuity between comparative and non-comparative work exists because their respective goals are identical: to explain social phenomena by establishing controls over the conditions and causes of variation.
- Any technique that furthers the goal of explaining variation is a comparative method. This includes virtually all analytic methods used by social scientists.

The Experiment

The ideal social scientific comparison is identical in structure to the simple experiment.

- The experiment, as a research design, facilitates establishing and quantifying cause-effect relationships between variables.
- It involves comparing two groups identical in all aspects except the experimental treatment given to one group.
- In this setup, only the treatment factor varies, while all other conditions remain constant or are randomized.
- Should significant post-treatment differences emerge between the experimental and control groups, these are ascribed to the treatment variable, thereby suggesting a tentative cause-effect sequence.

Randomization and Random Sampling

- *Randomization* in experiments (the process of assigning units to different groups in an experiment by chance) ensures that conditions are equalized across different groups, thereby isolating the effect of the variable being tested.
- Randomization is not to be confunded with *random sampling*.
  - *Randomization* starts with the *already chosen sample*, and divide it into groups using random assignment.
  - Randomization is essential for ensuring *internal validity* in experiments by equalizing conditions across groups.
  - *Random sampling* ensure *representativeness* of the sample, thereby facilitating generalization of the findings to the broader population.

Statistical Isolation and Regression Analysis

Statistical techniques like Multiple Regression Analysis can help to isolate the effect of one variable while controlling all the others.

- Often an experiment is neither possible nor recommendable (for example, in terms of costs or ethics), but the same comparative logic underlying the experiment is expressed in other research designs.
- Techniques like regression analysis are used in observational studies to control for confounding variables. They isolating the effect of the variables of interest by *holding all other variables constant.*

Statistical Isolation and Regression Analysis (example)

I use the following R code to simulate a data set of an observational research (for example, a survey) where the subjects are defined by the variables of age, gender (control variables), the independent variable of interest is the level of boredom, and the variable dependent is the time of using screens.

```r
set.seed(123)  # For reproducibility

# Number of observations
n <- 200

# Simulating independent variables
age <- rnorm(n, mean=30, sd=10)
gender <- sample(c("Male", "Female"), n, replace=TRUE)
boredom <- runif(n, min=1, max=10)  # Boredom scores between 1 and 10

# Simulating dependent variable (screen_time)
screen_time <- 5 + 0.05 * age - 0.3 * as.numeric(gender == "Female") + 0.4 * boredom +
screen_time <- pmin(pmax(screen_time, 1), 10)  # Restricting values to be within 1 and

```

Statistical Isolation and Regression Analysis (example)

Exploring the data set is a fundamental step to understand your data

```
# Viewing the first few rows of the dataset
head(dataSet)
```

```
       Age Gender  Boredom ScreenTime
1 24.39524   Male 3.135067   7.400233
2 27.69823   Male 7.178413   8.087625
3 45.58708   Male 3.032366   7.857552
4 30.70508   Male 3.866451   8.052993
5 31.29288   Male 2.565854   8.261682
6 47.15065   Male 8.212866   8.992132
```

Statistical Isolation and Regression Analysis (example)

Exploring the data set by calculating the summary statistics of variables is a fundamental step to understand your data

```r
# Calculate summary statistics
summary(dataSet$Boredom)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.011   3.217   5.429   5.442   7.797   9.954 
```

```r
summary(dataSet$ScreenTime)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  5.343   7.423   8.573   8.423   9.562  10.000 
```

```r
summary(dataSet$Age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  6.908  23.742  29.413  29.914  35.684  62.410 
```

```r
table(dataSet$Gender)
```

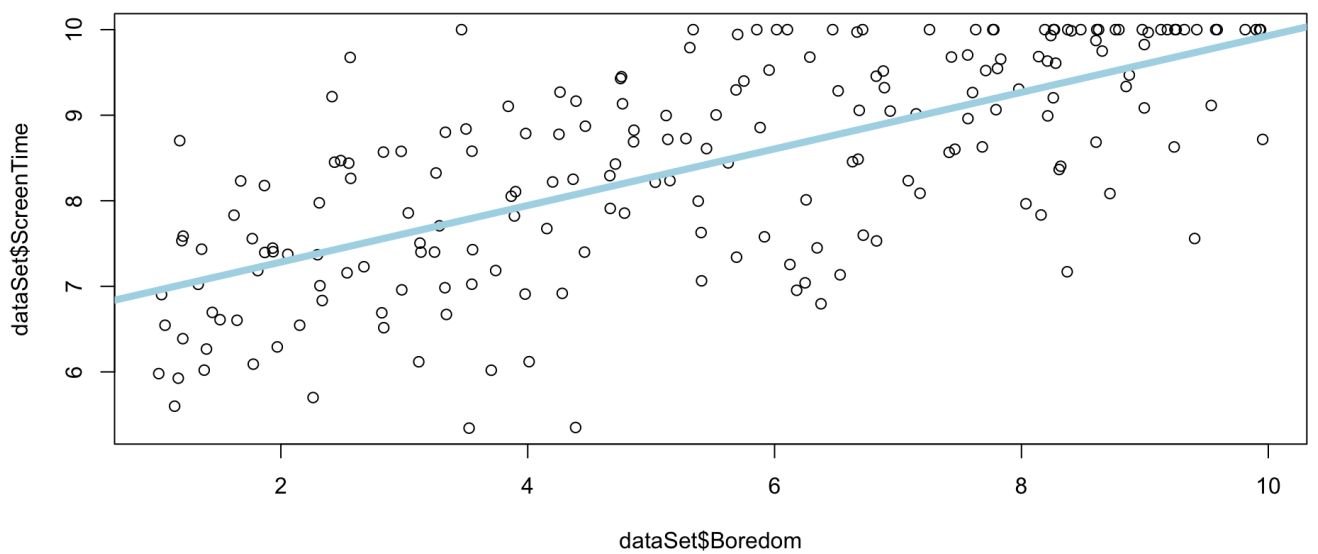
```
Female   Male 
   107     93 
```

Statistical Isolation and Regression Analysis (example)

Drawing a plot of the variables is a useful way to explore their relationship

```r
plot(dataSet$Boredom, dataSet$ScreenTime)

abline(lm(ScreenTime ~ Boredom, data = dataSet),
       col="lightblue", lwd = 5)
```



Statistical Isolation and Regression Analysis (example)

The Pearson Correlation Coefficient (r) gives a measure of the linear correlation between variables

```
1 cor.test(dataSet$Boredom, dataSet$ScreenTime)
```

```
    Pearson's product-moment correlation

data:  dataSet$Boredom and dataSet$ScreenTime
t = 13.466, df = 198, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6112877 0.7574723
sample estimates:
      cor
0.6913913
```

Statistical Isolation and Regression Analysis (example)

```r
reg_model <- lm(ScreenTime ~ Boredom + Age + Gender, data = dataSet)
summary(reg_model)
```

```
Call:
lm(formula = ScreenTime ~ Boredom + Age + Gender, data = dataSet)

Residuals:
     Min       1Q   Median       3Q      Max
-2.28479 -0.51265  0.09079  0.57258  1.95746

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.546412   0.254852  21.763  < 2e-16 ***
Boredom     0.338648   0.023111  14.653  < 2e-16 ***
Age         0.034476   0.006460   5.337  2.6e-07 ***
GenderMale  0.004508   0.121628   0.037     0.97
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.854 on 196 degrees of freedom
```

Most Similar and Most Different Research Designs

- John Stuart Mill's 1843 *"A System of Logic"* introduced comparison methods to identify causes in complex settings, leading to the development of two comparative research designs.
  - *Most Similar Systems Design (MSSD)*, based on his "method of difference," compares similar cases differing in the independent variable and outcome.
  - In the *Most Different Systems Design (MDSD)*, based on Mill's "method of agreement," cases vary in independent variables but share the same outcome, allowing the exclusion of these variables as causes of the outcome.

Most Similar and Most Different Research Designs

- MSSD and MDSD can be applied in both qualitative and quantitative research frameworks:
  - In qualitative research, the sample size is usually small, with the number of cases typically ranging from about 3 to 10 (Esser & Vliegenthart, 2017).
  - Quantitative research, even with few cases, uses datasets with many observations per case to enable statistical analysis.
- Case selection involves researcher discretion and knowledge of the cases, typically using purposive sampling.

Quasi-Experiments

- The logic of comparison also underpins quasi-experiments. This research design is different from experiments because it does not use random assignment of individuals to groups, but uses groups that already exist.
- The treatement can be either manipulated by the researcher or occurring by accident (for instance, a natural event or policy change). In the second case, the quasi-experiment are called *natural experiments*.

Pretest-Posttest Quasi-Equivalent Groups Design

- The pretest-posttest quasi-equivalent groups design is a quasi-experimental design that uses:
  - two or more experimental groups
  - a pretest and a posttest on those groups
  - does not use random assignment
  - and can either use a manipulated or natural IV.

Group 1	**O**	**X**	**O**
Group 2	**O**		**O**

- Example: Assess if training enhances patient-doctor communication confidence compared to no training. Group 1 is trained, Group 2 isn't; both evaluated before and after training

Interrupted Time Series Design

- Interrupted Time Series Design is a quasi-experimental research design that uses multiple measures of pretests and post-tests (pre-intervention and post-intervention).

Group 1 O O O X O O O

- Example: Does training boost patients' confidence in doctor interactions? The study involves one group taking four pretests, undergoing training, and then four posttests on perceived confidence.

Difference in Difference

- Difference in Difference is a quasi-experimental research design and the related statistical technique that:
  - employs a case and a control group, where a treatment affects one group but not another
  - compare the outcomes between these groups over time and particularly before and after the treatment
  - is based on the key assumption that the treatment and control groups would have followed similar trends over time in the absence of the treatment

Difference in Difference