

EDSD - Shapley Decomposition (Exercises) + Solutions

Preparation

Install the following package from GitHub:

```
install.packages("remotes")
remotes::install_github("elbersb/shapley")
```

Exercises – Part 1

1. Proof $\bar{y}_2 - \bar{y}_1 = \Delta p + \Delta y$.

Answer: Pull out the factor $\frac{1}{2}$ and rewrite as one sum. Expand all terms. The ‘mixed’ terms then cancel out and $\bar{y}_2 - \bar{y}_1$ remains.

2. How does the term Δy relate to standardization?

Answer: Δy is the age-standardized difference in the overall means, when we use $\frac{p_{i1} + p_{i2}}{2}$ as the reference distribution.

3. In R, use the `children` dataset to decompose the difference in overall rates using Kitagawa’s method.

```
children <- data.frame(
  age = c("20 to 24", "25 to 29", "30 to 34", "35 to 39", "40 to 44"),
  n1 = c(27, 152, 224, 239, 211),
  rate1 = c(37.037, 19.079, 15.179, 5.021, 6.161),
  n2 = c(363, 208, 96, 59, 48),
  rate2 = c(90.083, 76.923, 56.250, 20.339, 10.417)
)
```

Answer:

```
# create proportions
children$p1 <- children$n1 / sum(children$n1)
children$p2 <- children$n2 / sum(children$n2)

y1 <- with(children, sum(p1 * rate1))
y2 <- with(children, sum(p2 * rate2))
# we want to decompose
y2 - y1
```

```
[1] 60.60422
```

```
# two terms:
(delta_p <- with(children, sum((rate1 + rate2) / 2 * (p2 - p1))))
```

```
[1] 23.07169
```

```
(delta_y <- with(children, sum((p1 + p2) / 2 * (rate2 - rate1))))
```

```
[1] 37.53253
```

```
(delta_p + delta_y == y2 - y1)
```

```
[1] TRUE
```

Exercises - Part 2

1. What happens if the value function is additive in the factors? (e.g., let $N = \{a, b, c\}$, and $v(N) = a + b + c$, $v(\{a, c\}) = a + c$, $v(\emptyset) = 0$, etc.)

Answer: The contribution of factor a will simply be a , the contribution of factor b will simply be b , and so on. As $\sum_{S \subseteq N \setminus \{k\}} \frac{|S|!(m-1-|S|)!}{m!} = 1$ (see exercise below), the Shapley decomposition approach gives the desired answer.

Difficult take-home exercises (if you really want to!)

1. Proof $\sum_{S \subseteq N \setminus \{k\}} \frac{|S|!(m-1-|S|)!}{m!} = 1$. (This is to see that $\varphi_i(N, v)$ is truly a weighted average, with the weights summing to 1.)

Answer: Note that the only varying term in the sum is $|S|$, the size of the subset. If we have m players, the sum consists of $2^{(m-1)}$ terms. There will be always be one subset of size 0, $(m-1)$ subset(s) of size 1, $\binom{m-1}{2}$ subset(s) of size 2, and so on. In fact, the binomial coefficient gives the number of different subsets:

$$2^m = \sum_{k=0}^m \binom{m}{k} = \sum_{k=0}^m \frac{m!}{k!(m-k)!}.$$

Using this logic, we can rewrite the sum such that we don't iterate over all of the different subsets, but over the different sizes of subsets (i.e., from 0 to $m-1$), where we use $\binom{m-1}{k}$ as the weight:

$$\begin{aligned} \sum_{S \subseteq N \setminus \{k\}} \frac{|S|!(m-1-|S|)!}{m!} &= \sum_{k=0}^{m-1} \binom{m-1}{k} \frac{k!(m-1-k)!}{m!} \\ &= \sum_{k=0}^{m-1} \frac{(m-1)!}{k!(m-1-k)!} \frac{k!(m-1-k)!}{m!} \\ &= \sum_{k=0}^{m-1} \frac{(m-1)!}{m!} \\ &= m \frac{(m-1)!}{m!} \\ &= 1 \end{aligned}$$

This also shows that for the Shapley value, each set of subsets of the same size contribute equally.

2. Proof that $v(N) - v(\emptyset) = \varphi_1 + \varphi_2 + \dots + \varphi_m$. (This is called the efficiency property.)

Hint: The key to the proof is to write the Shapley value as a [telescoping series](#), such that the intermediary terms cancel out, and only the initial term, $v(N)$, and the final term, $v(\emptyset)$, remain.

Exercises - Part 3

1. Try the Shapley approach to Kitagawa's decomposition on your own. We defined $v(\emptyset) = \sum_i p_{i1} y_{i1}$ and $v(\{\Delta p, \Delta y\}) = \sum_i p_{i2} y_{i2}$. What happens when you switch these definitions?

2. Rewrite example 3 such that the value function uses a data frame. Try the decomposition on the `austria_chile` data provided below (data from 1981, fertility is per 1,000).

```
austria_chile <- data.frame(  
  factor = c("prop_unmarried", "nonm_fertility", "m_fertility"),  
  `austria` = c(0.41952, 23.99823, 71.83691),  
  `chile` = c(0.475, 50.82674, 115.73732)  
)
```

Answer:

```
library(shapley)  
  
example3 <- data.frame(  
  factor = c("prop_unmarried", "nonm_fertility", "m_fertility"),  
  `year_1963` = c(0.296, 0.011, 0.139),  
  `year_1983` = c(0.417, 0.019, 0.095)  
)  
  
il <- function(factors, data) {  
  # this assumes that the rows and columns are in the correct order!  
  alpha <- ifelse("prop_unmarried" %in% factors, data[1, 3], data[1, 2])  
  beta <- ifelse("nonm_fertility" %in% factors, data[2, 3], data[2, 2])  
  gamma <- ifelse("m_fertility" %in% factors, data[3, 3], data[3, 2])  
  alpha * beta / ((alpha * beta) + (1 - alpha) * gamma)  
}  
  
shapley(il, c("prop_unmarried", "nonm_fertility", "m_fertility"),  
  silent = TRUE, data = example3)
```

	factor	value
1	prop_unmarried	0.03377972
2	nonm_fertility	0.03470956
3	m_fertility	0.02445887

And the new data:

```
shapley(il, c("prop_unmarried", "nonm_fertility", "m_fertility"),  
  silent = TRUE, data = austria_chile)
```

	factor	value
1	prop_unmarried	0.04028401
2	nonm_fertility	0.13487862
3	m_fertility	-0.08529177

3. Bongaarts (1978) expressed the total fertility rate (TFR) as the product of five factors: the total fecundity rate, and the indices of proportion married, noncontraception, induced abortion, and lactational infecundability. The dataset `bongaarts` contains data for South Korea, 1960 and 1970. Use the Shapley value approach to decompose the difference in the TFR between 1960 and 1970 into the contributions of the five factors.

```
bongaarts <- data.frame(
  factor = c("fecundity", "marriage", "noncontracept", "abortion", "lactation"),
  `1960` = c(16.158, 0.720, 0.970, 0.970, 0.560),
  `1970` = c(16.573, 0.580, 0.760, 0.840, 0.660)
)
```

Answer:

```
library("shapley")

tfr <- function(factors = c()) {
  v <- bongaarts[["X1960"]]
  for (i in 1:nrow(bongaarts)) {
    if (bongaarts[[i, "factor"]] %in% factors)
      v[i] <- bongaarts[[i, "X1970"]]
  }
  prod(v)
}

tfr()
```

```
[1] 6.129875
```

```
tfr() == prod(bongaarts[["X1960"]])
```

```
[1] TRUE
```

```
tfr(c("fecundity", "marriage", "noncontracept", "abortion", "lactation"))
```

```
[1] 4.050102
```

```
tfr(c("fecundity", "marriage", "noncontracept", "abortion", "lactation")) == prod(bongaarts[
```

```
[1] TRUE
```

```
shapley(tfr, c("marriage", "noncontracept", "abortion", "lactation", "fecundity"),  
        silent = TRUE)
```

	factor	value
1	marriage	-1.0909785
2	noncontracept	-1.2298438
3	abortion	-0.7278399
4	lactation	0.8398707
5	fecundity	0.1290187

Additional exercises

1. Decompose the R^2 of a regression model into the contributions of each variable.

Answer: Check example 2 [here](#).

2. [Das Gupta \(1993\)](#) provides a survey of many decomposition procedures used in demography. Pick a decomposition procedure for which he provides data and see whether the Shapley decomposition yields identical results.
3. The dataset `income` below provides information for five individuals on their market income, the tax they pay, the transfers they receive, and their final income (final income = market income + tax + transfer). The Gini index is an inequality measure: if everyone has the same income, the Gini is 0. If one person has all the income, the Gini approaches 1. Taxes and transfers often reduce the Gini, in this case by about 17%. Decompose this difference into the contribution of taxes and transfers.

```
# 100 persons, each has 1 => Gini 0  
ineq::Gini(rep(100, 1))
```

```
[1] 0
```

```
# one person has 100, 100 persons have zero => Gini of 0.99  
ineq::Gini(c(100, rep(0, 100)))
```

```
[1] 0.990099
```

```
income <- data.frame(  
  MarketIncome = c(1, 20, 30, 40, 50),  
  Tax1 = c(0, -1, -2, -3, -4),  
  Tax2 = c(-5, -4, -3, -2, -1),  
  Transfer = c(9, 7, 5, 3, 1),  
  FinalIncome = c(5, 22, 30, 38, 46))  
  
ineq::Gini(income[["MarketIncome"]])
```

```
[1] 0.3347518
```

```
ineq::Gini(income[["FinalIncome"]])
```

```
[1] 0.2780142
```

Answer: Check example 3 [here](#).