Appendix E. The Shapley-Owen-Shorrocks Decomposition

Given an arbitrary function $Y = f(X_1, X_2, ..., X_n)$, the Shapley-Owen-Shorrocks decomposition is a method to decompose the value of $f(\cdot)$ into each of its arguments $X_1, X_2, ..., X_n$. Intuitively, the contribution of each argument if it were to be *"removed"* from the function. However, because the function can be nonlinear the order in which the arguments are removed matters in general for the decomposition. The function f can be the outcome of a regression, like the predicted values or sum of square residuals, or the output of a structural model, such as a counterfactual value for a variable given a list of model parameters or components, or a transformation of the sample, for example the Gini coefficient.

The Shapley-Owen-Shorrocks decomposition is the unique decomposition satisfying two important properties. First, the decomposition is exact decomposition under addition, letting C_j denote the contribution of argument X_j to the value of the function $f(\cdot)$,

$$\sum_{j=1}^{n} C_j = f(X_1, X_2, ..., X_n), \tag{E.1}$$

so that $C_j/f(\cdot)$ can be interpreted as the proportion of $f(.)$ that can be attributed to X_j.[39] Second, the decomposition is symmetric with respect to the order of the arguments. That is, the order in which the variable X_j is removed from $f(\cdot)$ does not alter the value of C_j.

The decomposition that satisfies both those properties is

$$C_j = \sum_{k=0}^{n-1} \frac{(n-k-1)!k!}{n!} \left(\sum_{s \subseteq S_k \setminus \{X_j\} : |s|=k} \left[f(s \cup X_j) - f(s) \right] \right), \tag{E.2}$$

where n is the total number of arguments in the original function f, $S_k \setminus \{X_j\}$ is the set of all *"sub-models"* that contain k arguments and exclude argument X_j.[40] For example,

$$\begin{aligned} S_{n-1} \setminus X_n &= f(X_1, X_2, ..., X_{n-1}) \\ S_1 \setminus X_n &= \{f(X_1), f(X_2), ..., f(X_{n-1})\}. \end{aligned}$$

[39] The interpretation holds as long as f is non-negative. If f can take negative values, then the interpretation of C_j under the exact additive rule can be misleading as some arguments can have $C_j < 0$.

[40] We abuse notation here. A sub-model is an evaluation of function f with only some of its arguments. This language is motivated by the function corresponding in practice to the outcome of a regression or structural model. Formally when we write $f(X_1)$ we mean $f(X_1, \emptyset_2, ..., \emptyset_n)$, where we assume the j-th argument of the function can always take on a null value denoted $\emptyset_j$. In our regression example below this null value corresponds to a zero valued regressor or parameter. In the case of structural model this null value can corresponding to setting some parameters to a predetermined value or excluding certain model components, like the adjustment of prices or a specific shock agents face.

The decomposition in (E.2) accounts for all possible permutations of the decomposition order. Thus, $\frac{(n-k-1)!k!}{n!}$ can be interpreted as the probability that one of the particular sub-model with k variables is randomly selected when all model sizes are all equally likely. For example, if $n = 3$, there are sub-models of size $\{0, 1, 2\}$. In particular, there are 2^2 permutation of models that exclude each variable: $\{\underbrace{(0,0)}_{k=0}, \underbrace{(1,0),(0,1)}_{k=1}, \underbrace{(1,1)}_{k=2}\}$.

$$
\begin{aligned}
k = 0 : \frac{(n-k-1)!k!}{n!} &= \frac{(3-0-1)!0!}{3!} = \frac{1}{3} \\
k = 1 : \frac{(n-k-1)!k!}{n!} &= \frac{(3-1-1)!1!}{3!} = \frac{1}{6} \\
k = 2 : \frac{(n-k-1)!k!}{n!} &= \frac{(3-2-1)!2!}{3!} = \frac{1}{3}
\end{aligned}
$$

Non-linear example

We illustrate the value of this decomposition with a simple non-linear model including $n = 3$ variables:

$$Y = f(X_1, X_2, X_3) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 X_2. \tag{E.3}$$

The objective is to decompose the value of Y into the contribution (or partial effect) of each variable.

Removing X_1

There are 4 possible models that exclude X_1, one with no variable, 2 with one variable and one with 2 variables

$$
\begin{aligned}
k = 0 &: \beta_0 \\
k = 1 &: \{\beta_0 + \beta_2 X_2\ ,\ \beta_0\} \\
k = 2 &: \beta_0 + \beta_2 X_2 + \beta_3 X_3 X_2
\end{aligned}
$$

In all 4 models, the partial effect of including X_1 is always $f(s \cup X_1) - f(s) = \beta_1 X_1$. This reflects the fact that the order that the order in which variables are included does not matter to construct C_1:

$$C_1 = \sum_{k=0}^{2} \frac{(3-k-1)!k!}{3!} \left(\sum_{s \subseteq S_k \backslash \{X_3\} : |s| = k} \left[f(s \cup X_j) - f(s) \right] \right) = \beta_1 X_1 \tag{E.4}$$

This would be the same for any argument X_j entering linearly into f an arbitrary number of variables: $Y = f(X_1, X_2, X_3, X_4, ..., X_n) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 X_2 + \sum_{j=4}^{n} \beta_j X_j$. The only difference is that the number of sub-models grows exponentially, 2^{n-1}, but the partial effect of including X_j for some $j \in \{4, \ldots, n\}$ is always $C_j = \beta_j X_j$.

Removing X_2

In this case, the partial effect can be decomposed into all the possible ways X_2 can be added into the model, $f(s \cup X_2) - f(s)$, these are

$$\begin{aligned}
k = 0\ (\emptyset_1, \emptyset_3) &: \beta_0 + \beta_2 X_2 - \beta_0 \;=\; \beta_2 X_2 \\
k = 1\ (X_1, \emptyset_3) &: \beta_0 + \beta_1 X_1 + \beta_2 X_2 - (\beta_0 + \beta_1 X_1) \;=\; \beta_2 X_2 \\
k = 1\ (\emptyset_1, X_3) &: \beta_0 + \beta_2 X_2 + \beta_3 X_2 X_3 - \beta_0 \;=\; \beta_2 X_2 + \beta_3 X_2 X_3 \\
k = 2\ (X_1, X3) &: \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 X_3 - (\beta_0 + \beta_1 X_1) \;=\; \beta_2 X_2 + \beta_3 X_2 X_3
\end{aligned}$$

Here, the partial effects of adding X_2 are not the same across sub-models because X_2 enters non-linearly into the original model. The symmetric property of the decomposition takes care of this.

$$\begin{aligned}
C_2 &= \underbrace{\frac{1}{3}\beta_2 X_2}_{k=0} + \underbrace{\frac{1}{6}(\beta_2 X_2) + \frac{1}{6}(\beta_2 X_2 + \beta_3 X_2 X_3)}_{k=1} + \underbrace{\frac{1}{3}(\beta_2 X_2 + \beta_3 X_2 X_3)}_{k=2} \\
&= \beta_2 X_2 + \frac{1}{2}\beta_3 X_2 X_3
\end{aligned} \tag{E.5}$$

The result is quite intuitive. $\beta_2 X_2$ appears in all sub-models, hence its probability of appearing in the decomposition is 1. $\beta_3 X_2 X_3$ appears in 2 of the 4 sub-models, hence its probability of appearing is 1/2. Weighting each term by its probability of appearing in the decomposition ensures symmetry.

Removing X_3

We proceed in the same way for X_3 as we did for X_2. There are 4 sub-models. In 2 of them the effect of adding X_3 is null because X_2 is not in the model. In the 2 remaining sub-models the effect is $\beta_3 X_2 X_3$. Hence,

$$C_3 = \frac{1}{2}\beta_3 X_2 X_3. \tag{E.6}$$

Finally, we verify the decomposition:

$$\begin{aligned}
C_1 + C_2 + C_3 &= \beta_1 X_1 + \left(\beta_2 X_2 + \frac{1}{2}\beta_3 X_2 X_3\right) + \left(\frac{1}{2}\beta_3 X_2 X_3\right) \\
&= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 X_3 \\
&= f(X_1, X_2, X_3) - \beta_0 \\
&= f(X_1, X_2, X_3) - f(\emptyset_1, \emptyset_2, \emptyset_3).
\end{aligned}$$

Note: The decomposition is additive with respect to the reference "null" model where none of the variables are included. This is made apparent in the previous result, where the decomposition does not include the value of β_0.

R-Squared

Finally, we consider a decomposition of the coefficient of determination in the linear model. Our use of the decomposition applies this for a non-linear model (combining the insights from this and the preceding example).

Consider a linear regression model with n regressors and $i = 1, \ldots, M$ observations,

$$y_i = \mathbf{x}_i'\beta + u_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} + u_i, \tag{E.7}$$

and define the average value of y as $\overline{y} \equiv \sum_{i=1}^{M} y_i/M$ and the predicted value

$$\hat{y}_i = \mathbf{x}_i'\hat{\beta} = \hat{\beta}_0 + \sum_{j=1}^{n} \hat{\beta}_j x_{ij}, \tag{E.8}$$

where we assume that all regressors have zero mean so that $\hat{\beta}_0 = \overline{y}$.

The function of interest is $f(X_1, \ldots, X_K) = R^2$, defined as the explained sum of squares SSE over the total sum of squares SST

$$R^2(X_1, X_2, .., X_n) = \frac{SSE}{SST} = \frac{\sum_{i=1}^{M}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{M}(y_i - \overline{y})^2}. \tag{E.9}$$

This makes it clear that the function being decomposed is non-linear even though the model that generates it is itself linear.

Note: The reference value for the R^2 in the Shapley-Owen-Shorrocks decomposition is given by the model without regressors, satisfying

$$R^2(\emptyset) = \frac{\sum_i^M(\hat{\beta}_0 - \overline{y})^2}{\sum_i^M(y_i - \overline{y})^2} = 0, \tag{E.10}$$

so that, in this case, the decomposition recovers the level of the R^2 of the full model (with all variables), unlike the previous example.

Details of the decomposition when n = 3 Consistent with the previous example, we show the decomposition for $n = 3$ regressors. As before, we abuse notation by only listing the arguments being included in each sub-model. The contribution of each variable is:

$$\begin{aligned} R_1^2 = {} & \frac{1}{3}\left[R^2(X_1) - R^2(\emptyset)\right] + \frac{1}{6}\left(\left[R^2(X_1, X_2) - R^2(X_2)\right] + \left[R^2(X_1, X_3) - R^2(X_3)\right]\right) \\ & + \frac{1}{3}\left[R^2(X_1, X_2, X_3) - R^2(X_2, X_3)\right]; \end{aligned} \tag{E.11}$$

$$
\begin{aligned}
R_2^2 = {} & \frac{1}{3}\left[R^2(X_2) - R^2(\emptyset)\right] + \frac{1}{6}\left(\left[R^2(X_1, X_2) - R^2(X_1)\right] + \left[R^2(X_2, X_3) - R^2(X_3)\right]\right) \\
& + \frac{1}{3}\left[R^2(X_1, X_2, X_3) - R^2(X_1, X_3)\right]; \qquad \text{(E.12)} \\
R_3^2 = {} & \frac{1}{3}\left[R^2(X_3) - R^2(\emptyset)\right] + \frac{1}{6}\left(\left[R^2(X_3, X_2) - R^2(X_2)\right] + \left[R^2(X_1, X_3) - R^2(X_1)\right]\right) \\
& + \frac{1}{3}\left[R^2(X_1, X_2, X_3) - R^2(X_2, X_1)\right]. \qquad \text{(E.13)}
\end{aligned}
$$

Summing across all the contributions we obtain back $R^2(X_1, X_2, X_3)$,

$$
R_1^2 + R_2^2 + R_3^2 = R^2 = f(X_1, X_2, X_3). \qquad \text{(E.14)}
$$

Note: The value of the contribution differs from the standard definition of partial R-squared. This is because the partial R-squared is an all else equal comparison of excluding regressor X_j from the regression. It does not satisfy the exact decomposition requirement, nor (when applied iteratively) the symmetry requirement.