Interpreting the Effects of Categorical Predictors

Jim Grace

This module considers the interpretation of path coefficients when modeling with categorical predictors.

This module follows the one entitled: “SEM Essentials – Interpreting Path Coefficients”, which should be studied first.

A general citation for this material is

Notes: IP-064929; Support provided by the USGS Climate & Land Use R&D and Ecosystems Programs. I would like to acknowledge formal review of this material by Gaoue Orou, University of Hawaii and James Cronin, U.S. Geological Survey. Thanks also to Tamara Ticktin, University of Hawaii and Elisabeth Brouwers, USGS for helpful comments. Any use of trade, form, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.
Last revised 17.01.31.
Source: https://www.usgs.gov/centers/wetland-and-aquatic-research-
Scientists often use standardized coefficients for interpretation (here I am referring to the classical method of standardizing based on standard deviations). This is helpful for putting all the path coefficients in the same units. However, when categorical predictors are involved, the interpretation of standardized coefficients becomes distorted. Here I show an easy way to address this problem. Along the way we peel back the cover on coefficients in general.

Note: Here I only illustrate the situation where we have categorical predictors that are binary (0,1) or Yes/No. Sometimes variables can have more than two states and are classified as “ordered categorical”, e.g., “Low, Medium, High”. In such a case, there are two choices. First (and most general) is the option of converting your single variable with three states into three dummy variables, Low (0,1); Medium (0,1); and High (0,1). You would then include two of the three variables in your model. One dummy variable must be omitted from the model to avoid singularity. The omitted state becomes the baseline against which the others are compared. So, if you omitted Low, then the tests for Medium and High are tests for whether responses for those levels are greater than for the Low class. Second approach is to treat the effects of your ordered categorical predictor as linear and then you can simply allow it to have values of 0, 1, or 2. Now there is a single coefficient and we
assume going from 0 to 2 is double that from 0 to 1.
The data for this illustration are extracted from a study that included the doubling of atmospheric CO$_2$.

Reference for this work is:


Note, this article was featured in Nature News April 9, 2009, featured in Nature Climate Change Research Highlights May 5, 2009, and was a USGS Science Newsroom Pick.

Here I use a “net-effect” model to illustrate the principle.

The net effect was a greater ability of marsh sods to build soil elevation under elevated CO$_2$.

A box plot gives some sense of the span of values relative to the mean response to CO$_2$ treatment.
"Reduced-form" is a common term in the SEM literature for models that capture net effects while omitting at least one, but sometimes many mediating nodes.

The original model was more complex than this and included mediating pathways. Here I show a “reduced-form” model that absorbs the full causal network into a net or total effect.
The data are simple, but the interpretation is particular.

View of the data*,
- 60 pots total
- CO₂ treatment (0,1)
- ElevChange (mm)

*These data can be found in the notes section of this slide.

Data for example if .csv file not available (semi-colons are end of line markers):

```
pot,CO2,ElevChange;
1,1,3.88141026; 2,1,1.33653846; 3,1,4.69230769; 4,1,18.3910256; 5,1,44.0769231; 6,1,2.99038462; 7,1,0.46153846; 8,1,28.1538462; 9,1,-2.3846154; 10,1,12.2307692; 11,1,1.33653846; 12,1,18.3910256; 13,1,50.7307692; 14,1,1.9230769; 15,1,-0.8076923; 16,0,19.6538462; 17,0,-4.5769231; 18,0,7.06153846; 19,0,-1.0384615; 20,0,1.07692308; 21,0,-1.3461538; 22,0,1.80769231; 23,0,6.38461538; 24,0,25.9230769; 25,0,-1.8461538; 26,0,4.04230769; 27,0,0.05448718; 28,0,25.9230769; 29,0,4.30769231; 30,0,4.80769231; 31,1,-7; 32,1,7.61538462; 33,1,19.5; 34,1,8.11538462; 35,1,0.15384615; 36,1,26.9020979; 37,1,25.5153846; 38,1,0.76923077; 39,1,31.2307692; 40,1,0.11538462; 41,1,21.6538462; 42,1,37.7307692; 43,1,8.30769231; 44,1,5; 45,1,5.80769231; 46,0,3.4775641; 47,0,-3.7692308; 48,0,31.7692308
```

Data from

Here I assume basic familiarity with lavaan. If you need a refresher, refer to the tutorial entitled “Introduction to lavaan”.

```r
# specify model
mod <- 'ElevChange ~ CO2'

# fit model
mod.fit <- sem(mod, data=dat)

# request output
summary(mod.fit, rsq=T, standardized=T)
```
Results, showing standardized and unstandardized coefficients.

```
\texttt{lavaan (0.5-15) converged normally after 1 iteration}

\begin{verbatim}
\begin{tabular}{llllll}
Number of observations & 60 \\
Estimator             & ML \\
Minimum Function Test Statistic & 0.000 \\
Degrees of freedom    & 0 \\
P-value (Chi-square)   & 1.000 \\

Estimate & Std.err & Z-value & P(>|z|) & Std.lv & Std.all \\
\hline
ElevChange ~ CO2 \textsuperscript{mean diff between CO2 treatments} & 5.280 & 3.701 & 1.427 & 0.154 & 5.280 & 0.181 \\
\hline
Variances:  \\
ElevChange & 205.457 & 37.511 & \textsuperscript{Std.all uses the std.dev of CO2} 205.457 & 0.967 \\
\hline
R-Square: \\
ElevChange & 0.033 \\
\end{verbatim}
```

The raw “Estimate” has a straightforward interpretation, the standardized relies on the std.dev of a categorical variable.

One should already be familiar with the difference between raw and standardized coefficients. Note that in lavaan, it prints two kinds of standardized coefficients, “Std.lv” and “Std.all”; the latter of these is what we want.

The raw coefficient/estimate here is 5.280. Its interpretation is explained on the next page.
Some might be tempted to log-transform elevation change because of its distribution. However, we are interested in interpreting the coefficients in original units and there is no biological reason to interpret the process of sediment building in log scale, so we will not.

So, what is the problem with interpreting standardized coefficients based on categorical predictors?

Raw estimate (5.280) is the mean different between the treatments (in elevation units, mm).

This is straightforward to interpret, but would be hard to compare to other path coefficients that are in different units.

I provide a refresher on the relationship between raw and standardized parameters on the next page.
Remember, standardized parameters are in standard deviation units.

```r
### Compute standardized coefficient by hand
est = 5.280
sd.elev <- sd(ElevChange)
sd.co2 <- sd(CO2)

std.all <- est*(sd.co2/sd.elev)
print(std.all)

> print(std.all)
[1] 0.181134
```

Here we reconstruct the standardized coefficient reported two slides ago.

So, standardized coefficients are predicted changes in units of standard deviations (predicted sd change in $y$ as function of sd change in $x$).
There has been a lot of opposition to standardized coefficients from some statisticians. Scientists must find some way to move forward, nonetheless, which is why classical standardization is so popular.

I propose that the interpretability of standardized coefficients depends on the fact that there is a relationship between standard deviations and ranges.

Generally, 6 standard deviations = 99% of the range for a true Gaussian distribution.

So, we can think of standardized coefficients as similar* to predicted changes in $y$ along its range as you vary $x$ along its range.

*Note that this only holds strictly for idealized Gaussian variables.
The relationship between standard deviations and ranges does not hold consistently for categorical variables.

```r
### What is sd of CO2 in this case?
print(sd.co2)

> print(sd.co2)
[1] 0.5042195

### What if we had a categorical variable with unequal numbers of 0s and 1s?
new.cat <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1)
print(sd(new.cat))

> print(sd(new.cat))
[1] 0.3077935
```

For case of equal numbers of 0s and 1s, then std.dev = 0.5. Certainly not the case that 6 std.dev = 1 range, as is assumed for Gaussian variables.

The standard deviation of a categorical variable does not have the same meaning as that of a normal variable. Since the range of categoricals is fixed at 1, the relationship between std dev and range varies based on the frequency of 0s and 1s. – Not helpful!
There is a useful alternative to conventional standardized coefficients – range standardization (Grace and Bollen 2005).

Range standardization provides a good option in this situation. (see tutorial “SEM Essentials - Interpreting Coefficients”)

```r
### Range standardization
range.elev <- max(ElevChange) - min(ElevChange)
range.co2 = 1

std.range <- est*(range.co2/range.elev)
print(std.range)

> print(std.range)
[1] 0.09145903
```

Here we show that the predicted change in elevation is 9% of its range if we double CO₂.

Source for this method is

Historical note: This method was developed after studying Pedhazur’s book on statistics and his extended discussion of the problems of interpreting standardized coefficients.

When standardizing by ranges, we should confirm that the computed range for ElevChange is appropriate for interpretation.

```
# R code to visualize
plot(ElevChange, pch=16)
```

The distribution of values across the range is reasonably continuous, which supports our use of range standardization.

I generally refer to this methodology as “relevant range” standardization. The investigator needs to select the relevant range for application of the coefficient. This need extends to raw coefficients as well, though that is rarely discussed.

Note that the majority of values observed is in the lower end of the distribution because the distribution of treatment combinations, not because of non-linear response form.
Graphical representation is now.

Effect is now in units of “change in soil elevation across its range” when CO$_2$ is doubled. Can be compared among different pathways now.

We point out that this is a small amount and non-significant based on conventional criteria. When the impact of increasing CO$_2$ is examined fully, however, there is a significant interactive effect that is hidden in this net effect (see Cherry et al. 2009. for the full story).