USGS
science for a changing world

Interpreting the Effects of Categorical Predictors

Jim Grace

U.S. Department of the Interior
U.S. Geological Survey

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

Grace, J.B. 2006. Structural Equation Modeling and Natural Systems. Cambridge University Press. Cambridge, UK.

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-

center/science/quantitative-analysis-using-structural-equation

How do we interpret the effects of categorical predictors?

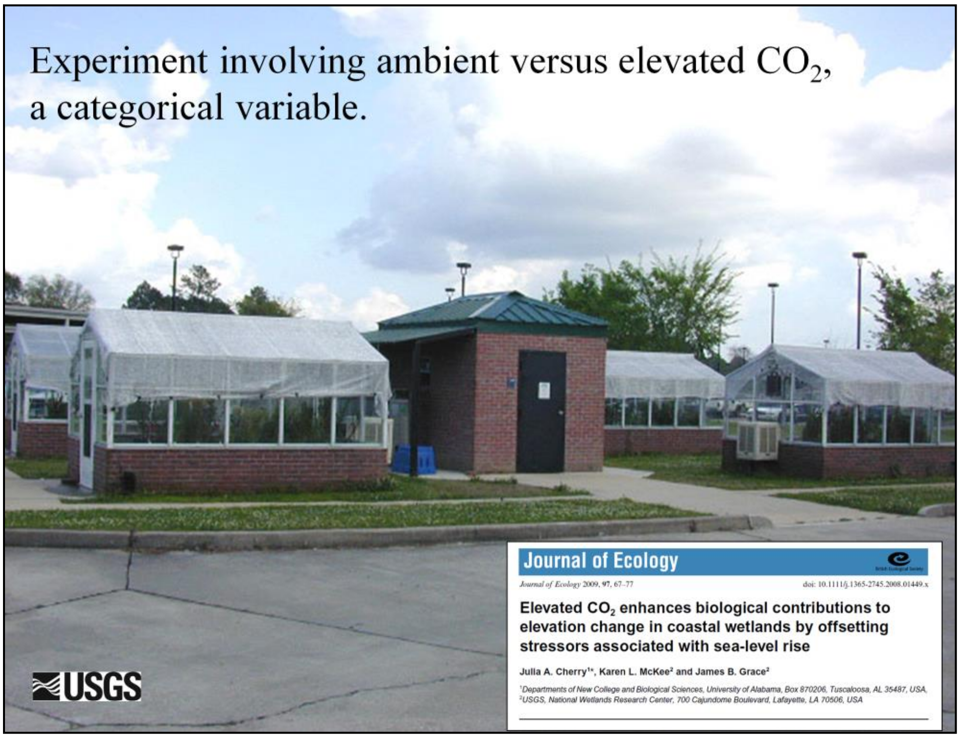
- Binary categorical predictors are often coded as (0,1) variables.
- No statistical problems with using categorical predictors. Assumptions about error distributions are associated with response variables only.
- However, there are some issues related to interpretation of categorical effects, illustrated here.
- Good time for "range standardization"!

It would probably be helpful if you review the module on "Interpreting Path Coefficients" before going through this module.

USGS 2

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.



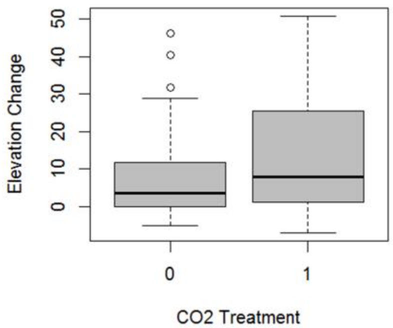
The data for this illustration are extracted from a study that included the doubling of atmospheric CO_2.

Reference for this work is:

Cherry, J.A., McKee, K.L., and Grace, J.B. 2009. Elevated CO_2 enhances biological contributions to elevation change in coastal wetlands by offsetting stressors associated with sea-level rise. *Journal of Ecology* 97:67-77.

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. http://www.nature.com/climate/2009/0905/full/climate.2009.32.html.

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



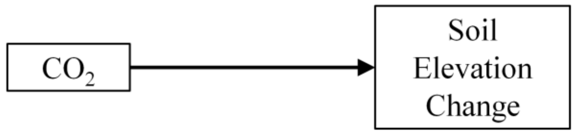
The net effect was a greater ability of marsh sods to build soil elevation under elevated CO_2.

USGS

4

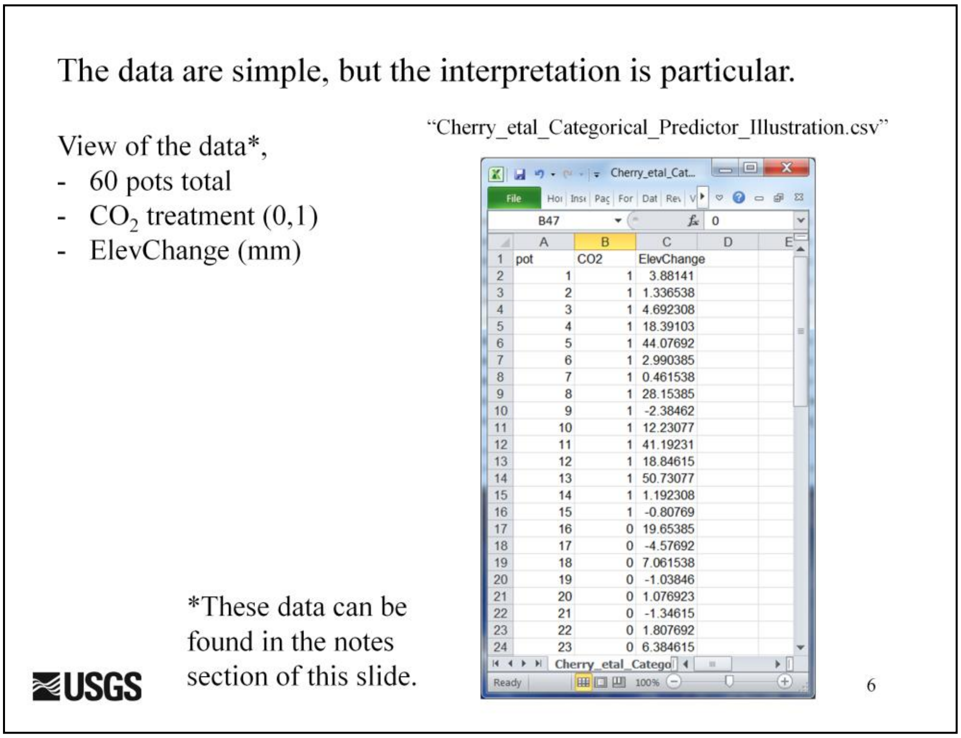
A box plot gives some sense of the span of values relative to the mean response to CO_2 treatment.

Graphical representation.



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.

“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.



	A	B	C
1	pot	CO2	ElevChange
2	1	1	3.88141
3	2	1	1.336538
4	3	1	4.692308
5	4	1	18.39103
6	5	1	44.07692
7	6	1	2.990385
8	7	1	0.461538
9	8	1	28.15385
10	9	1	-2.38462
11	10	1	12.23077
12	11	1	41.19231
13	12	1	18.84615
14	13	1	50.73077
15	14	1	1.192308
16	15	1	-0.80769
17	16	0	19.65385
18	17	0	-4.57692
19	18	0	7.061538
20	19	0	-1.03846
21	20	0	1.076923
22	21	0	-1.34615
23	22	0	1.807692
24	23	0	6.384615

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; 11,1,41.1923077; 12,1,18.8461538; 13,1,50.7307692; 14,1,1.19230769; 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; 26,0,40.4230769; 27,0,0.05448718; 28,0,28.8461538; 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;

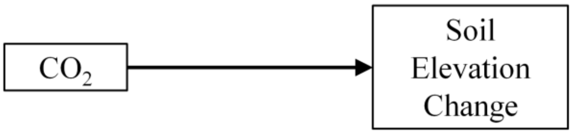
47,0,-3.7692308; 48,0,31.7692308

Data from

Cherry, J.A., McKee, K.L., and Grace, J.B. 2009. Elevated CO_2 enhances biological contributions to elevation change in coastal wetlands by offsetting stressors associated with sea-level rise. *Journal*

of Ecology 97:67-77.

lavaan coding is simple.




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

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

# request output
summary(mod.fit, rsq=T, standardized=T)
```

USGS

7

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

Results, showing standardized and unstandardized coefficients.

```
lavaan (0.5-15) converged normally after 1 iteration

  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
Regressions:
  ElevChange ~
    CO2               5.280    3.701    1.427    0.154    5.280   0.181

Variances:
    ElevChange 205.457   37.511                         205.457   0.967

R-Square:

    ElevChange        0.033
```

mean diff between CO2 treatments

Std.all uses the std.dev of CO2

USGS The raw “Estimate” has a straightforward interpretation, the standardized relies on the std.dev of a categorial variable. 8

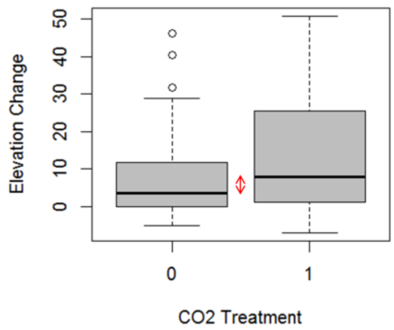
One should already be familiar with the difference between raw and standardized coefficients. Note that in lavaan, it prints two kinds of standardized coefficients, “Std.lv” and “Std.all”; the latter of these is what we want.

The raw coefficient/estimate here is 5.280. Its interpretation is explained on the next page.

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.

USGS

9

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.

Remember, standardized parameters are in standard deviation units.

```
### 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.

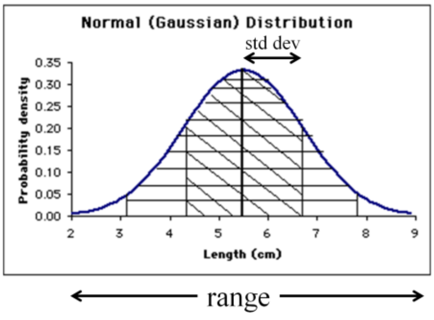
So, standardized coefficients are predicted changes in units of standard deviations (predicted sd change in *y* as function of sd change in *x*).

USGS

10

This material refers back to "SEM_1_6_Interpreting Coefficients".

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.

USGS

11

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.

The relationship between standard deviations and ranges does not hold consistently for categorical variables.

```
### 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(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1)
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.

USGS

12

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")

```
### 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 CO_2.

USGS

13

Source for this method is

Grace, J.B. and Bollen, K.A. 2005. Interpreting the results from multiple regression and structural equation models. Bulletin of the Ecological Society of America 86:283-295.

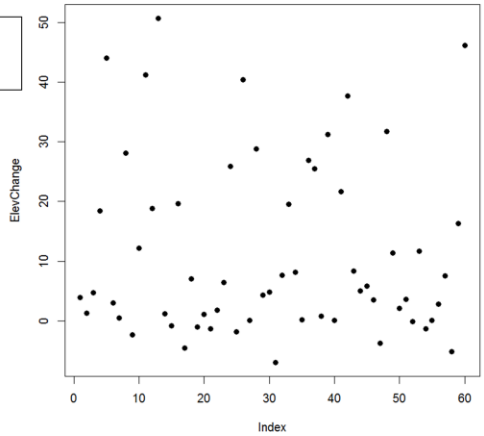
Historical note: This method was developed after studying Pedhazur's book on statistics and his extended discussion of the problems of interpreting standardized coefficients.

Pedhazur, E.J. 1997. Multple Regression in Behavioral Research. Wadsworth Publishing; 3 edition.

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.



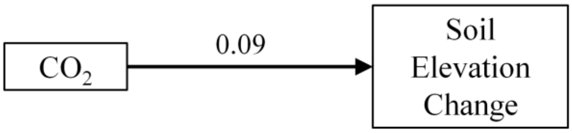
USGS

14

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).

USGS