Introduction to structural equation modeling and mixed models in 

Day 5 – Part 1: SEM

Oksana Buzhdygan

oksana.buzh@fu-berlin.de

Day 5 – Part 1

Outline

- Categorical Variables in SEM

Categorical Variables in SEM

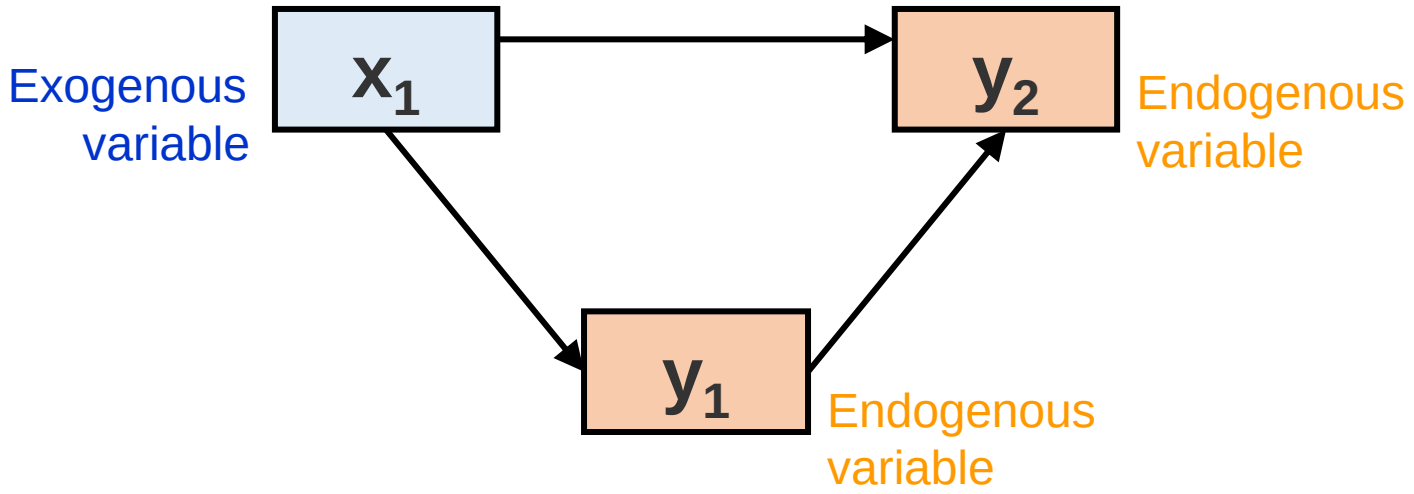
Categorical / discrete data

- binary (yes/no, failure/success, dead/alive, male/female),
- nominal (site 1, site 2, site 3)
- ordinal levels (small < medium < large; yang < middle < old).

Categorical Variables in SEM

Categorical / discrete data

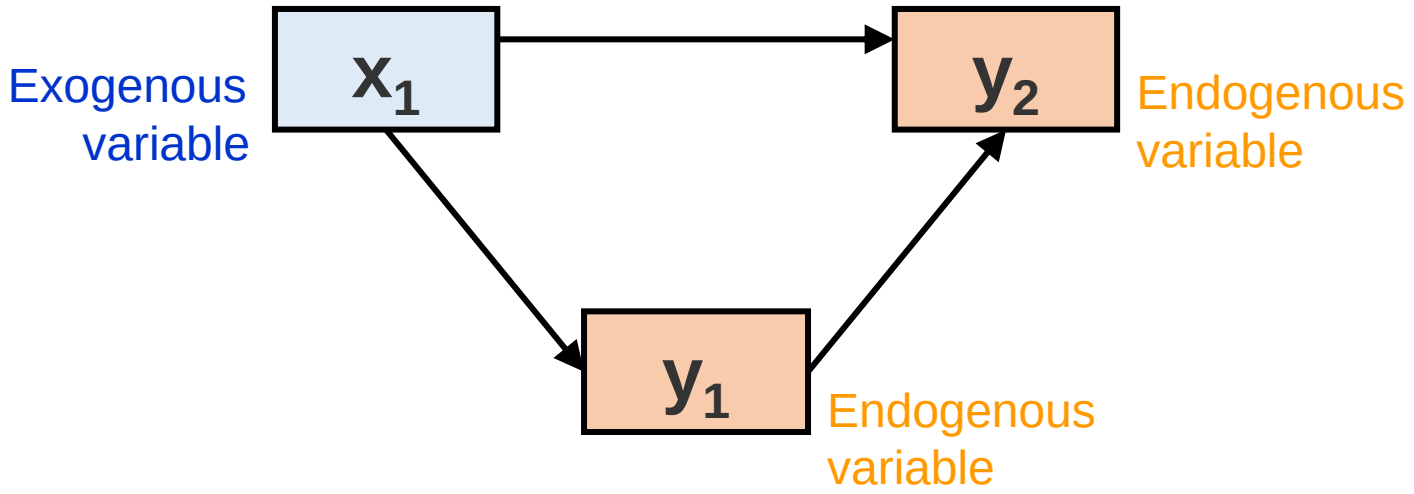
- binary (yes/no, failure/success, dead/alive, male/female),
- nominal (site 1, site 2, site 3)
- ordinal levels (small < medium < large; yang < middle < old).



Categorical Variables in SEM

Categorical / discrete data

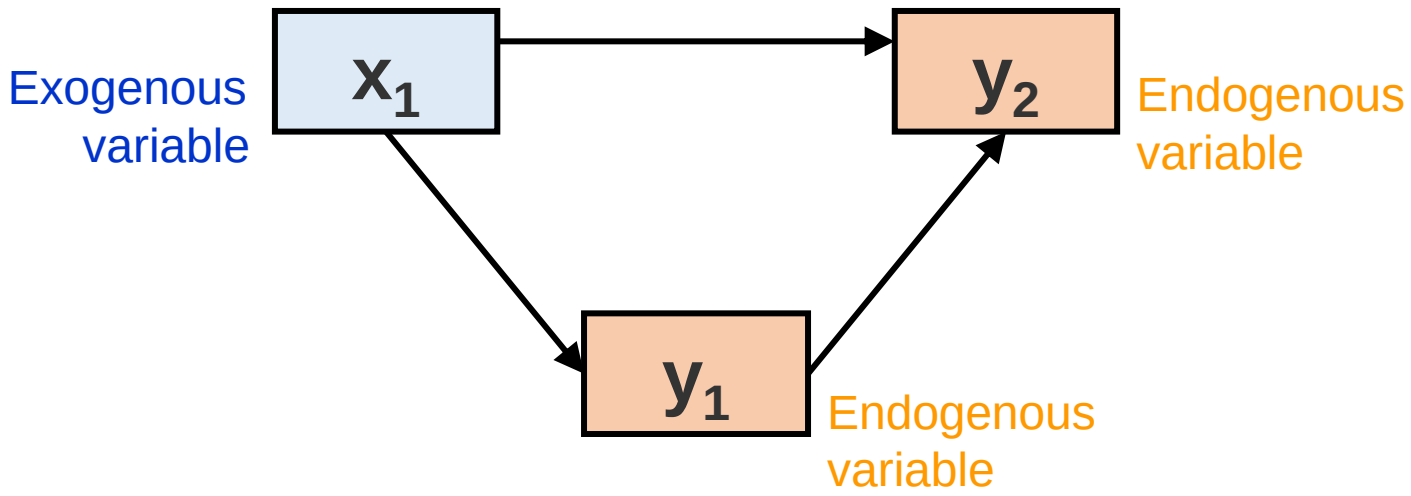
- binary (yes/no, failure/success, dead/alive, male/female),
- nominal (site 1, site 2, site 3)
- ordinal levels (small < medium < large; yang < middle < old).



Categorical Variables in SEM

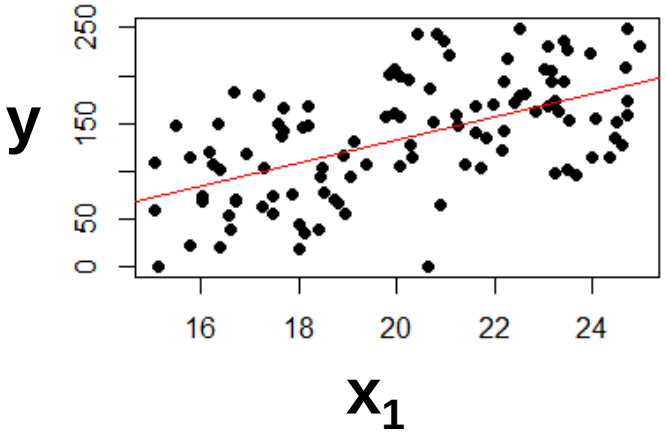
Categorical / discrete data

- binary (yes/no, failure/success, dead/alive, male/female),
- nominal (site 1, site 2, site 3)
- ordinal levels (small < medium < large; yang < middle < old).



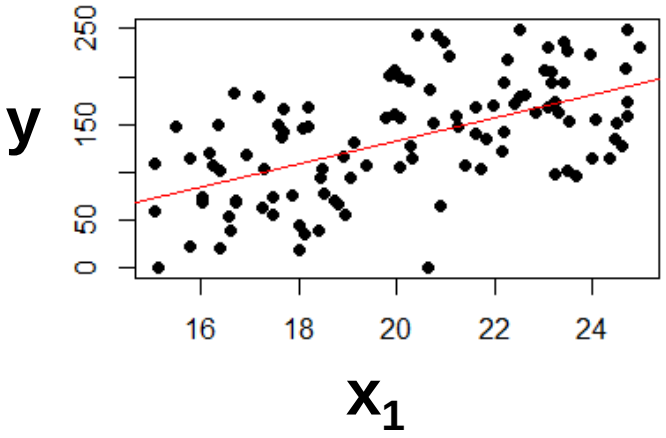
Categorical Variables in SEM

$$y = a + b_1 x_1 + \epsilon$$

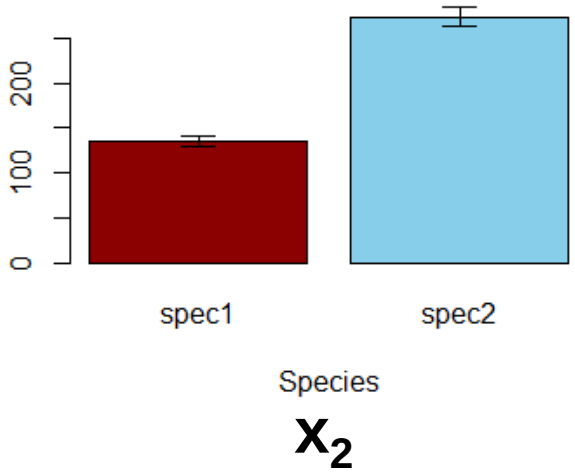


Categorical Variables in SEM

$$y = a + b_1 x_1 + \epsilon$$



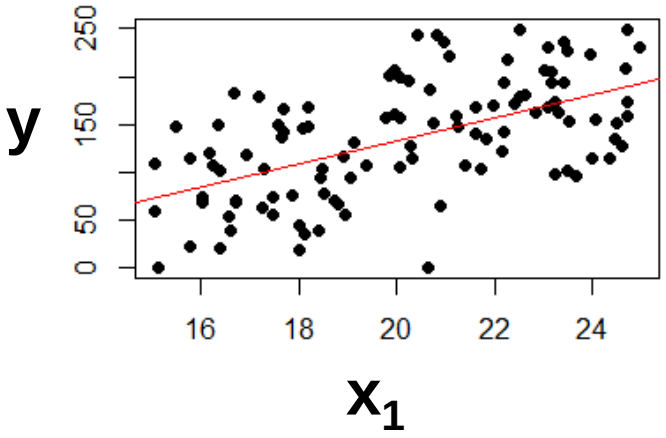
$$y = a + b_2 x_2 + \epsilon$$



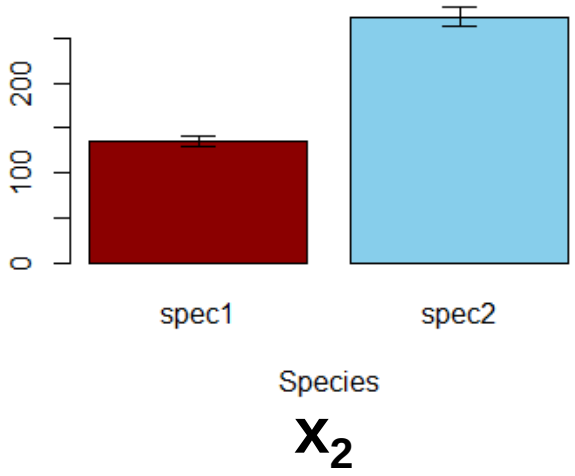
x_2
Species
spec1
spec1
spec2
spec1
spec2

Categorical Variables in SEM

$$y = a + b_1 x_1 + \epsilon$$



$$y = a + b_2 x_2 + \epsilon$$

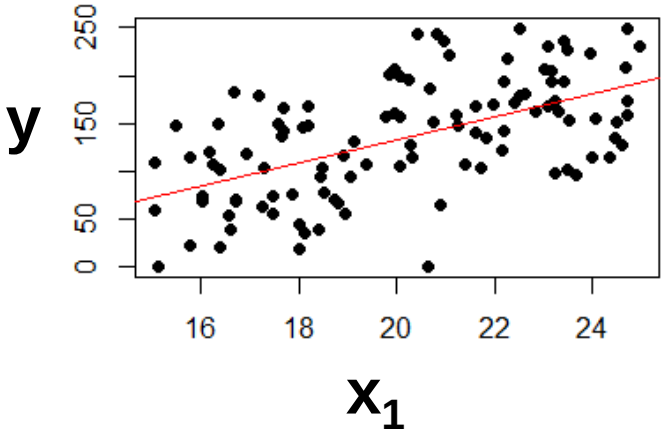


x_2

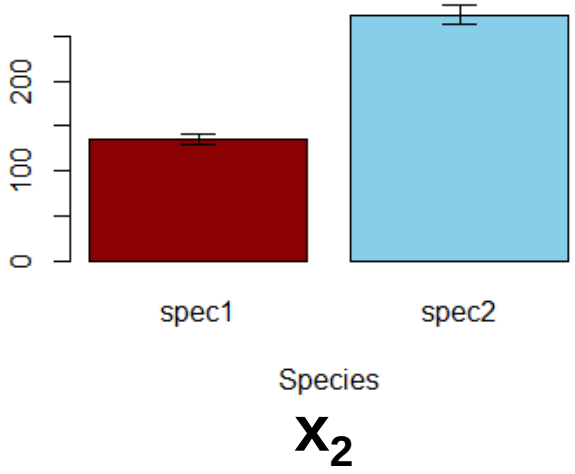
Species
spec1
spec1
spec2
spec1
spec2

Categorical Variables in SEM

$$y = a + b_1 x_1 + \epsilon$$



$$y = a + b_2 x_2 + \epsilon$$



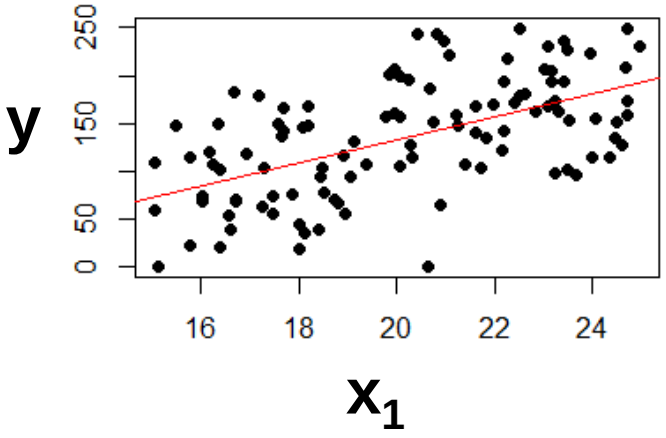
x_2

Species
spec1
spec1
spec2
spec1
spec2

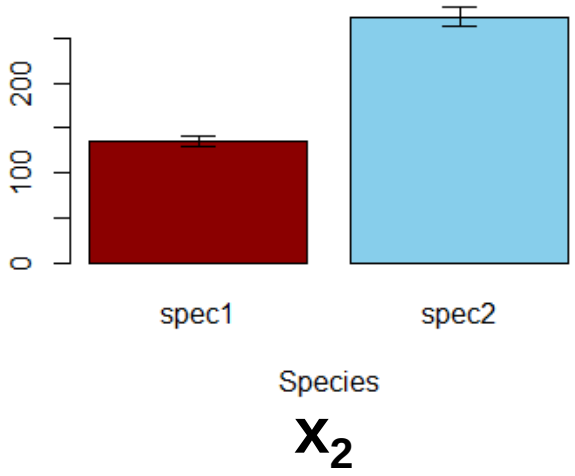
spec1	spec2
1	0
1	0
0	1
1	0
0	1

Categorical Variables in SEM

$$y = a + b_1 x_1 + \epsilon$$



$$y = a + b_2 x_2 + \epsilon$$



x_2

Species
spec1
spec1
spec2
spec1
spec2

spec1	spec2
1	0
1	0
0	1
1	0
0	1

Exogenous Categorical Variables

Approaches when we have Exogenous Categorical Variables:

1) for nominal, binary, or ordinal variables, create separate dummy variables for each factor levels (treat them as absent “0” or present “1”).
   - The key: for the factor with k levels use k-1 dummy variables (to avoid singularity)

2) for binary variables, set the values as 0 or 1 and model as numeric (yields a single coefficient).

3) for ordinal variables, set the values depending on the order of the factor, e.g., small = 1 < medium = 2 < large = 3, and then model as numeric (yields a single coefficient).

4) Use `piecewiseSEM`

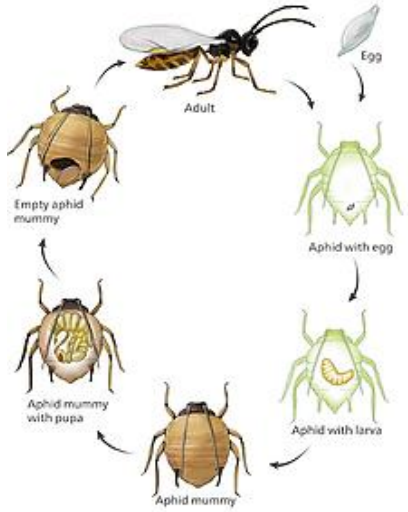
Example

Categorical Exogenous Variable

Biocontrol agents of crop-pests (aphids)

Lacewing larva

Parasitic wasp



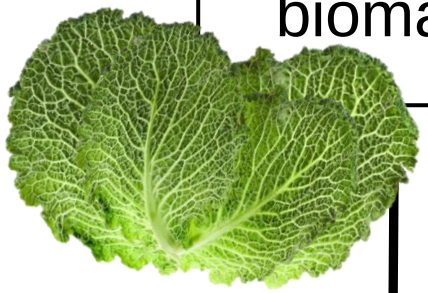
Intercropping with repellent plants

Example

Categorical Exogenous Variable

Garlic presence

Host-plant biomass



Aphid natural enemies:

No enemy

Predator present

Parasite present

Aphid density

150 experimental microcosms

Example

Categorical Exogenous Variable

```
# Read and check the data

aphid_data <- read.csv("Aphid_data.csv")

> str(aphid_data)

'data.frame': 150 obs. of  8 variables:

 $ aphid     : num  14.9 35.6 43.8 2.1 36.7 ...

 $ host_plant: num  38.8 40.7 46.9 35.2 50.9 ...

 $ garlic_ef : Factor w/ 2 levels "absent","present": 2 1 1 2 1 2 2 2 1 1 ...
 $ garlic    : int  1 0 0 1 0 1 1 1 0 0 ...

 $ enemy     : Factor w/ 3 levels "no", "predator", "parasite": 3 3 1 2 1 1 3 2 3 2 ...
 $ no_enemy  : int  0 0 1 0 1 1 0 0 0 0 ...
 $ predator  : int  1 1 0 0 0 0 1 0 1 0 ...
 $ parasite  : int  0 0 0 1 0 0 0 1 0 1 ...
```

binary variable

(0/1) - dummy variable for binary

nominal variable

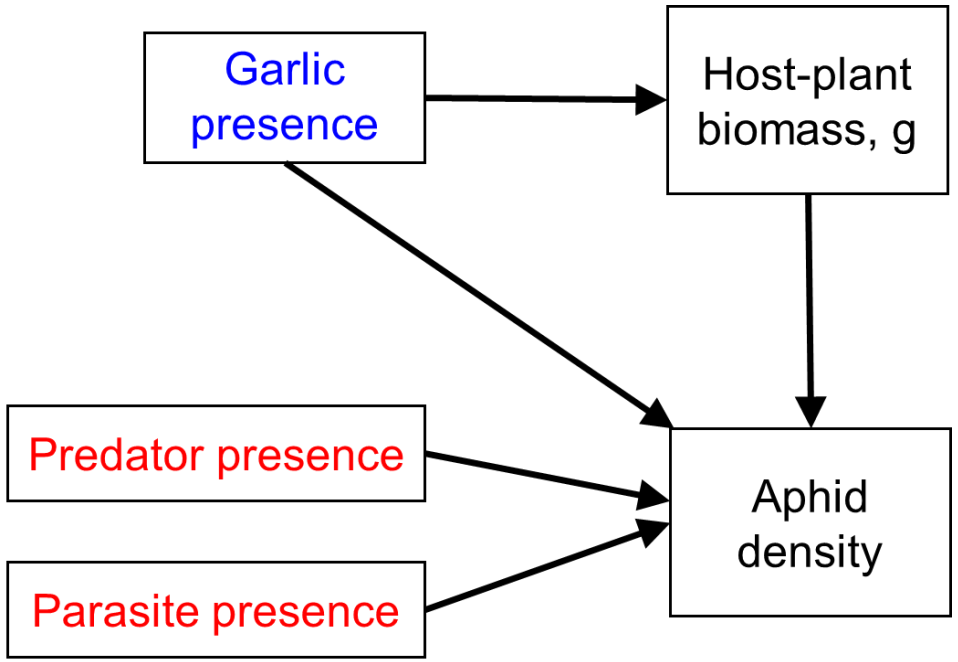
dummy variables created for each factor level

Example

Categorical Exogenous Variable

```
# specify and fit the model in lavaan

sem_mod <- ' aphid ~ host_plant + garlic + predator + parasite

             host_plant ~ garlic

'

fit <- sem(sem_mod, data=aphid_data)

summary(fit, standardize = T, rsq = T)
```

Only 2 out of 3 dummy variables are included



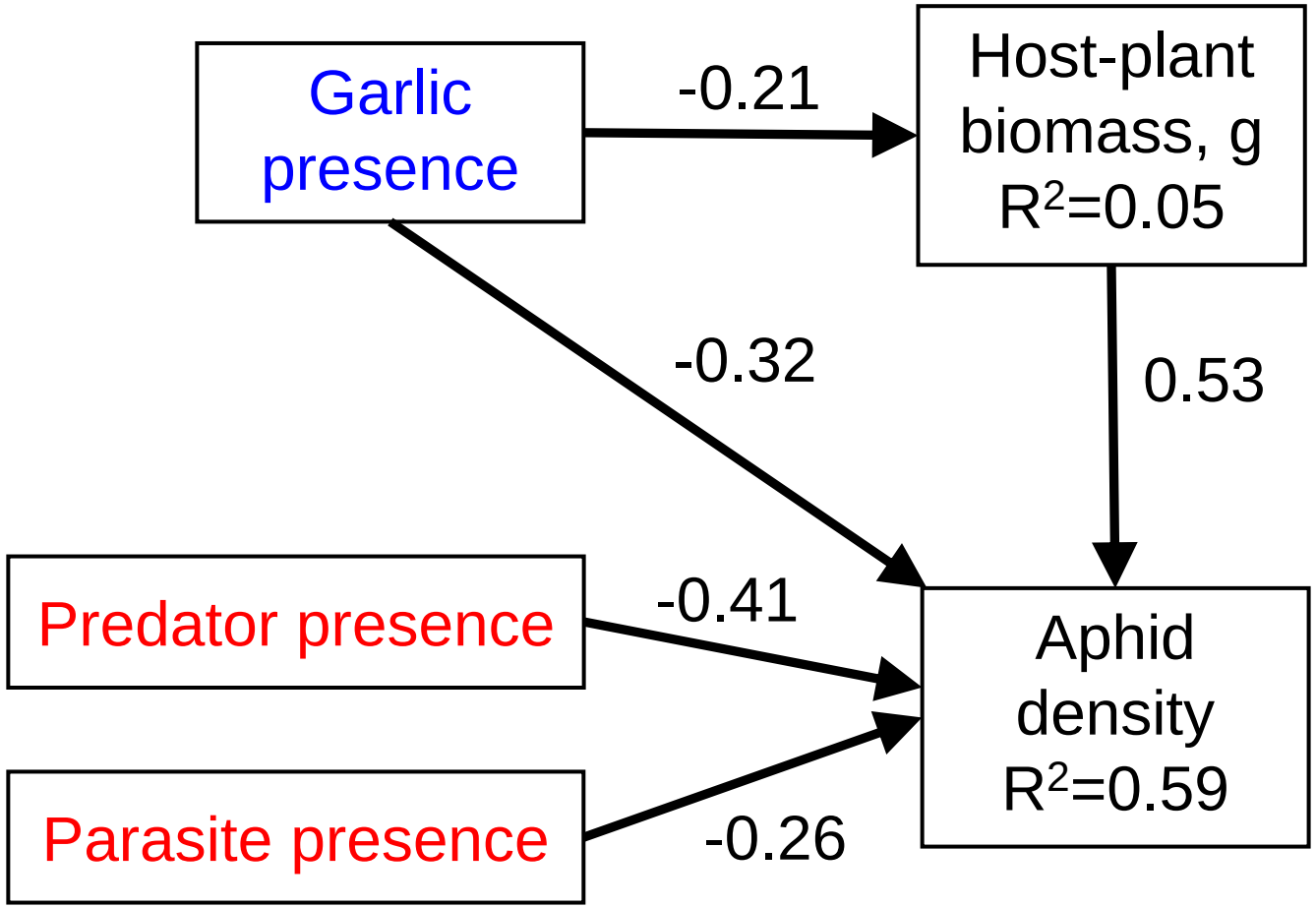
Example

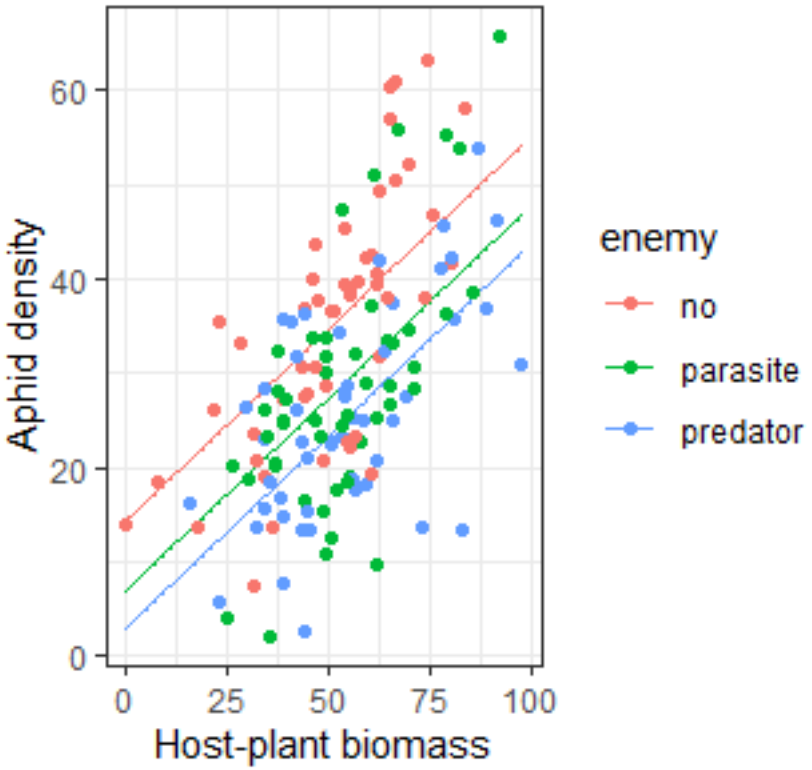
Categorical Exogenous Variable

```
# Results
Model Test User Model:
  Test statistic                                 1.658
  Degrees of freedom                                 2
  P-value (Chi-square)                           0.436

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  aphid ~
    host_plant        0.408    0.041    9.925    0.000    0.408    0.534
    garlic           -8.506    1.437   -5.921    0.000   -8.506   -0.321
    predator        -11.372    1.707   -6.663    0.000  -11.372   -0.405
    parasite         -7.375    1.712   -4.309    0.000   -7.375   -0.262
  host_plant ~
    garlic           -7.570    2.769   -2.734    0.006   -7.570   -0.218

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .aphid            72.753    8.401    8.660    0.000   72.753    0.414
   .host_plant      287.445   33.191    8.660    0.000  287.445    0.953

R-Square:
                   Estimate
    aphid             0.586
    host_plant        0.047
```

Example

Categorical Exogenous Variable



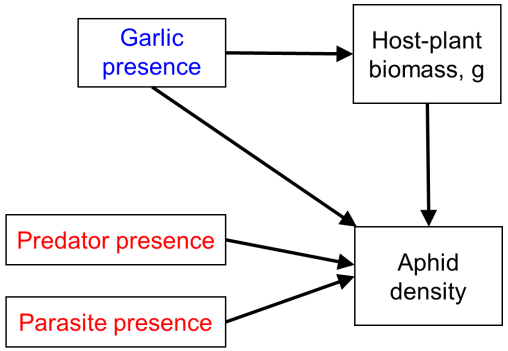


χ^2 = 1.65, DF=2, n=150, p = 0.43

Example

Categorical Exogenous Variable

```
# calculate indirect effects
sem_mod <- ' aphid ~ a1*host_plant + a2*garlic + predator + parasite
             host_plant ~ a3*garlic
                  # define indirect and total effect
                  direct := a2
                  indirect := a3*a1
                  total := direct + indirect
'
fit <- sem(sem_mod, data=aphid_data)
summary(fit, standardize = T, rsq = T, fit.measures=T)
>
Defined Parameters:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    direct           -8.506    1.437   -5.921    0.000   -8.506   -0.321
    indirect         -3.086    1.171   -2.636    0.008   -3.086   -0.116
    total           -11.592    1.800   -6.439    0.000  -11.592   -0.437
```



Example

Categorical Exogenous Variable

Garlic presence

-0.21

Host-plant biomass, g
R^2=0.05

-0.32

0.53

Aphid natural enemies

Predator presence

-0.41

Parasite presence

-0.26

Aphid density
R^2=0.59

Effects of **garlic presence** on **aphid density**:

direct: -0.32

indirect: -0.12

total: -0.44

χ^2 = 1.65, DF=2, n=150, p = 0.43 RMSEA=0, (CI = 0, 0.15) , p_{RMSEA}=0.55, CFI=1.00; SRMR=0.025

Endogenous Categorical Variables

Approaches when we have Endogenous Categorical Variables:

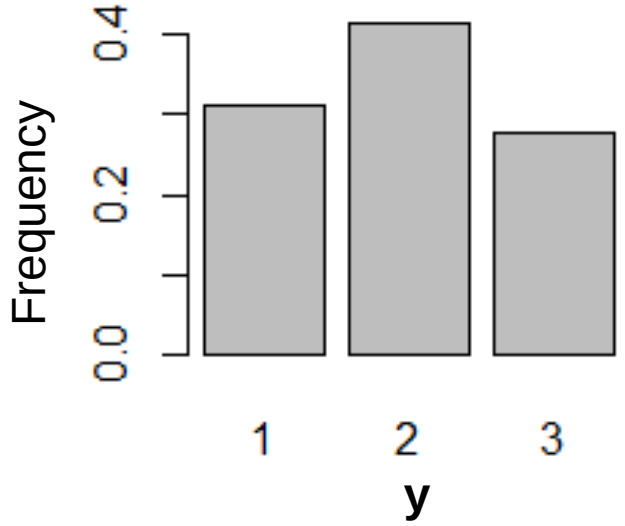
1) for binary and ordinal variables use the argument ‘ordered’ in ***lavaan*** with fitting function ‘sem’

2) for nominal variables (i.e., levels are not ordered) use the factor levels to construct a composite variable.

Endogenous Categorical Variables

- Normal distribution means continuous data
- Ordinal data can not be assumed normal

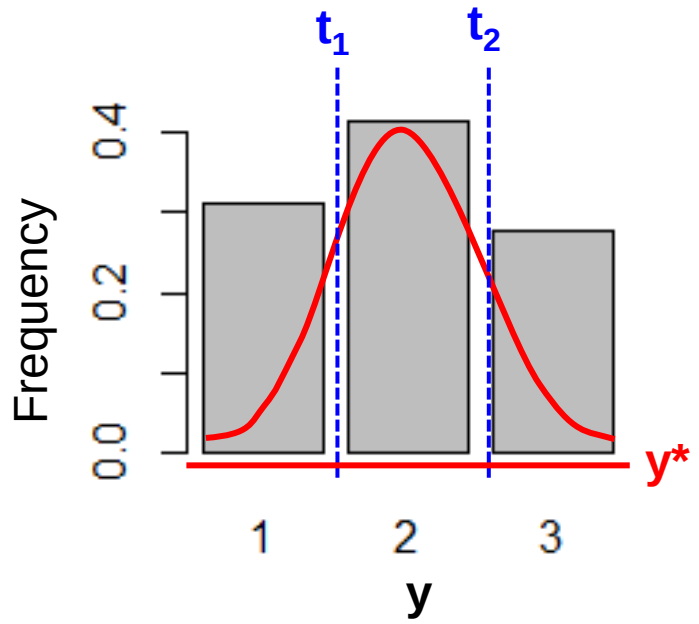
  Solution: to use the threshold models



Endogenous Categorical Variables

- Normal distribution means continuous data
- Ordinal data can not be assumed normal

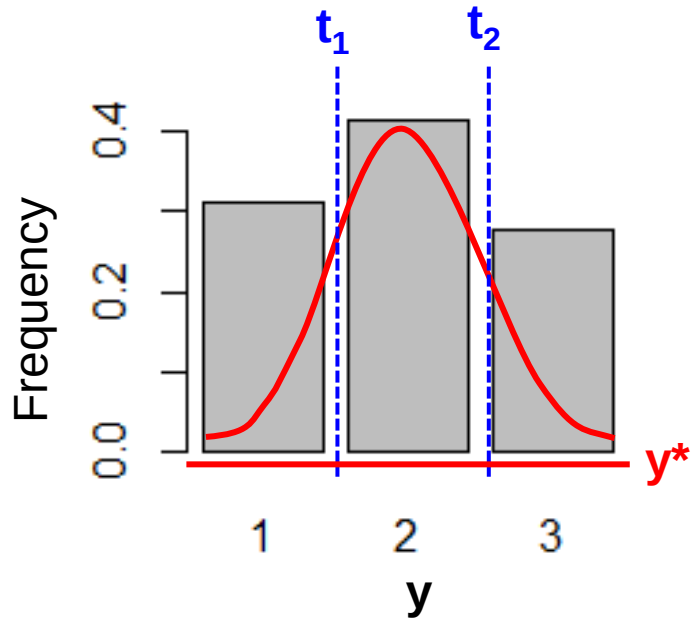
  Solution: to use the threshold models

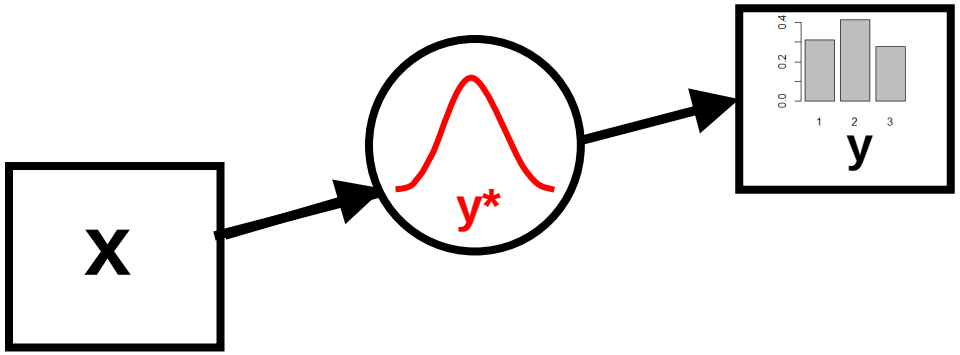


Endogenous Categorical Variables

- Normal distribution means continuous data
- Ordinal data can not be assumed normal

  Solution: to use the threshold models

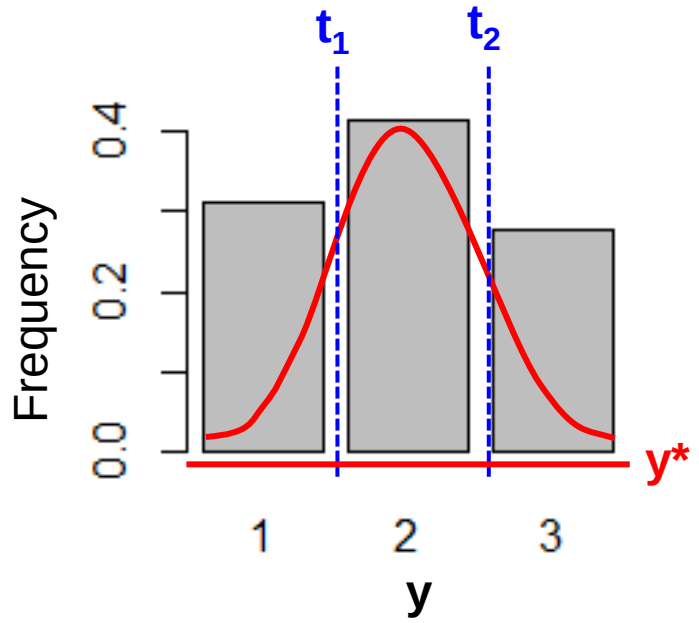




Endogenous Categorical Variables

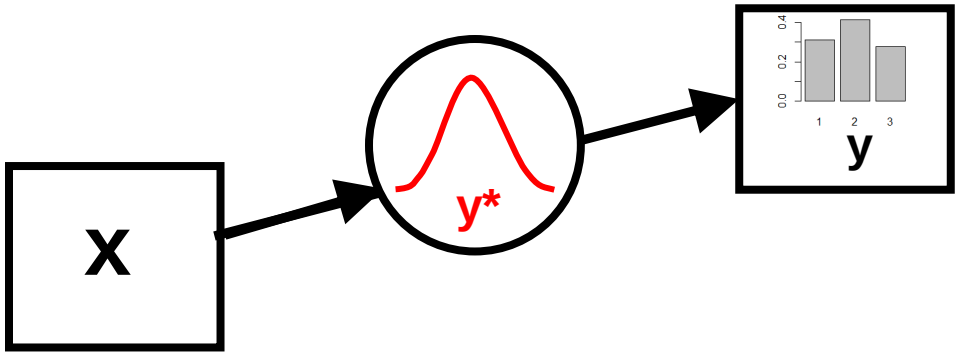
- Normal distribution means continuous data
- Ordinal data can not be assumed normal

  Solution: to use the threshold models



Estimation not via ML but via
(diagonally) weighted least squares (D)WLS

$$\boldsymbol{F_{WLS} = (s - \sigma)^{\top} \mathbf{W}^{-1} (s - \sigma)}$$



Example

Categorical Endogenous Variable

Human activities affect fish communities in ponds

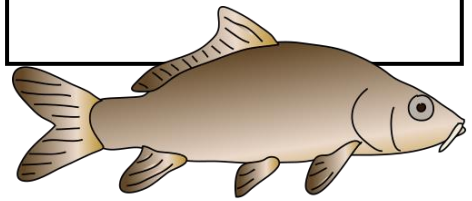
120 ponds

Example

Categorical Endogenous Variable

Human impact intensity

Presence of invasive fish



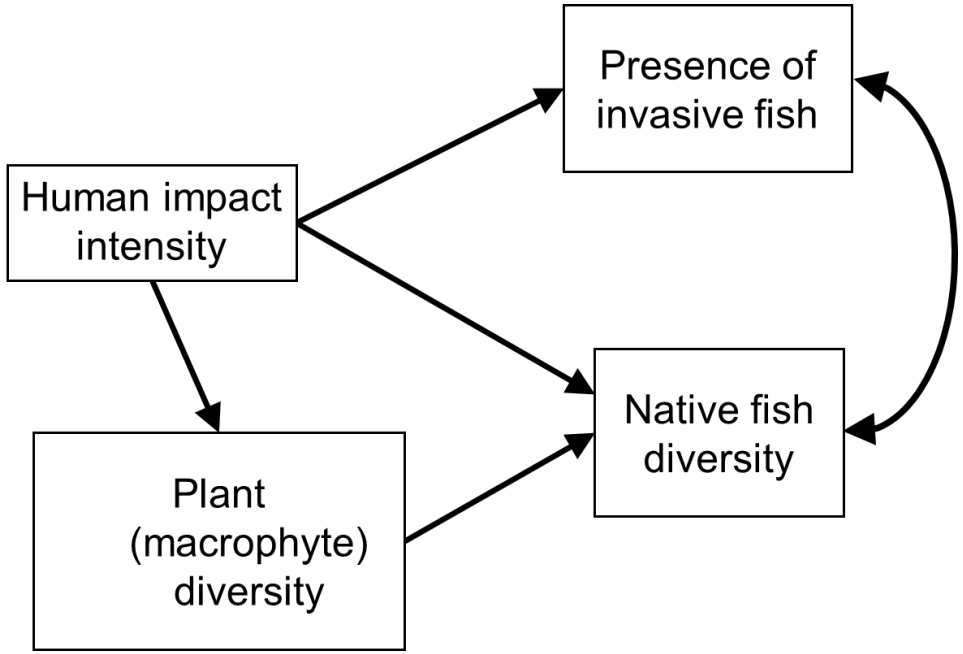
Native fish diversity

Plant (macrophyte) diversity

Example

Categorical Endogenous Variable

```
# Read and check the data
fish_data <- read.csv("Fish_data.csv")
str(fish_data)

sem_mod2 <- ' inv_fish ~  HII
               native_fish ~ plant_div + HII
               plant_div ~ HII
               native_fish ~~ inv_fish
'
fit2 <- sem(sem_mod2, data=fish_data,
                      ordered = c("inv_fish"))

summary(fit2, standardize = T, rsq = T)
```



Example

Categorical Endogenous Variable

```
# Read and check the data

Estimator                                      DWLS
  Optimization method                          NLMINB
  Number of model parameters                       10

  Number of observations                          120

Model Test User Model:
                                              Standard      Robust
  Test Statistic                                 0.022       0.022
  Degrees of freedom                                 1           1
  P-value (Chi-square)                           0.882       0.882
  Scaling correction factor                                  1.000
  Shift parameter                                            0.000
       simple second-order correction

Parameter Estimates:

  Standard errors                            Robust.sem
```

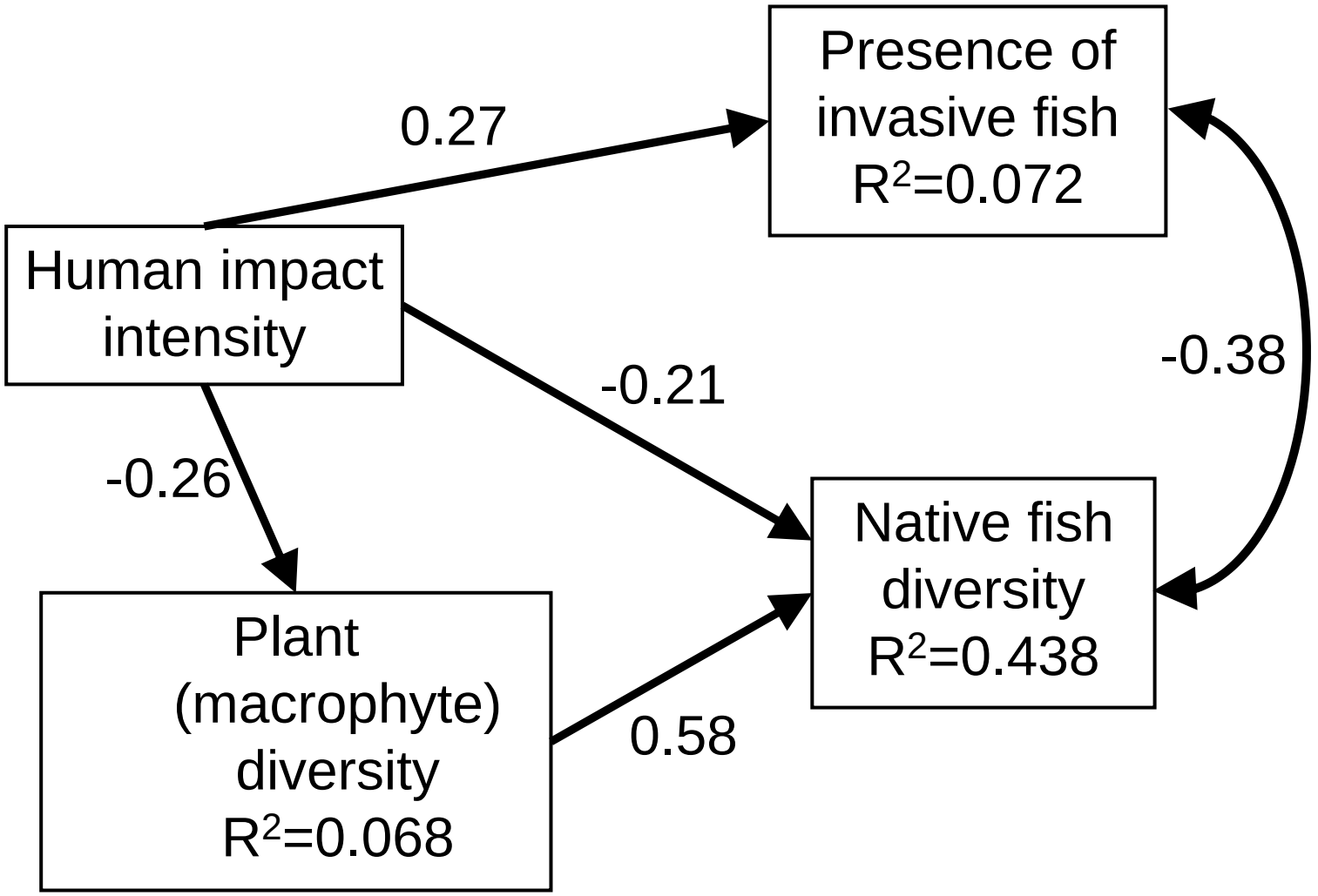
Example

Categorical Endogenous Variable

```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  inv_fish ~
    HII               0.308    0.128    2.411    0.016    0.308    0.268
  native_fish ~
    plant_div         0.475    0.059    7.994    0.000    0.475    0.576
    HII              -1.186    0.424   -2.797    0.005   -1.186   -0.210
  plant_div ~
    HII              -1.785    0.695   -2.569    0.010   -1.785   -0.261

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
 .inv_fish ~~
   .native_fish      -1.466    0.572   -2.561    0.010   -1.466   -0.383

Thresholds:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    inv_fish|t1       0.567    0.288    1.969    0.049    0.567    0.546

R-Square:
                   Estimate
    inv_fish          0.072
    native_fish       0.438
    plant_div         0.068
```

Example

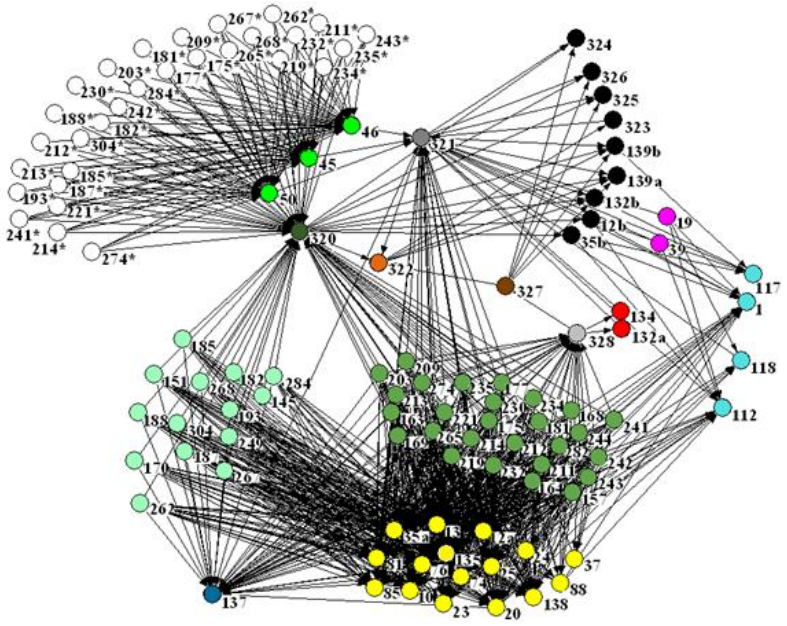
Categorical Exogenous Variable



χ^2 = 0.022, DF=1, n=120, p = 0.88

Day 5 Task 1

Effects of land use on arthropod food webs in grasslands



Food webs

235 grasslands

Food-web length

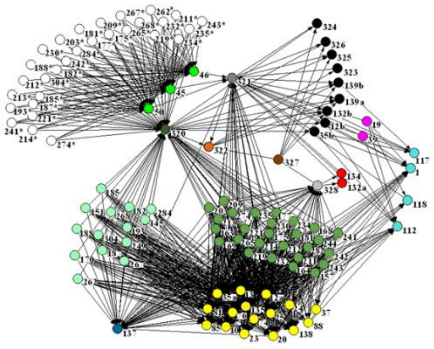
“1 level”: only herbivores and decomposers,
“2 levels”: carnivores present in addition to level 1,
“3 levels”: omnivores present in addition to level 1 and level 2.

Grazing intensity

(“low”, “medium”, or “high”)

Day 5 Task 1

Effects of land use on food webs in grasslands



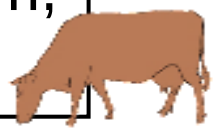
Food-web length
(1_level, 2_levels, 3_levels)

Plant biomass

Plant species richness

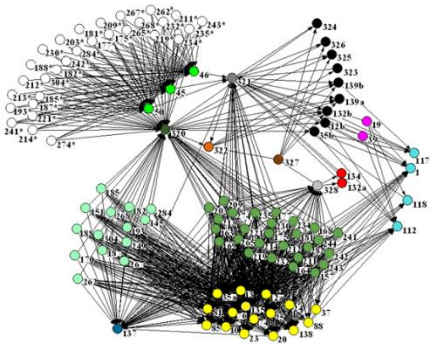
Soil organic C

Grazing
(low, medium, high)



Day 5 Task 1

Effects of land use on food webs in grasslands



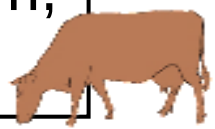
Food-web length
(1_level, 2_levels, 3_levels)

Plant biomass

Plant species richness

Soil organic C

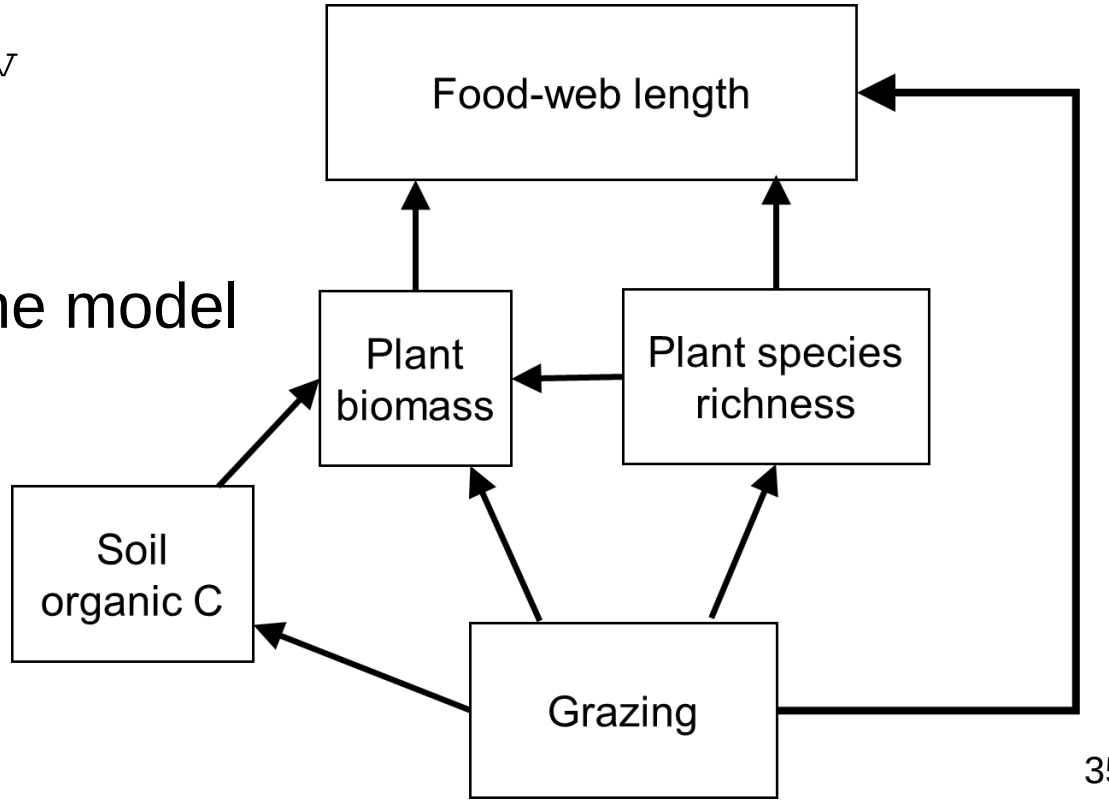
Grazing
(low, medium, high)



Day 5 Task 1

Effects of land use on food webs in grasslands

1. Specify the following model in lavaan
   - For this, if needed, recode the categorical variables in a way appropriate for the analysis
3. Fit the model using data `Food-web_data.csv`
4. Get the fit indices
5. Fill in Standardized Coeficients and R^2 for the model
6. Think about how to interpret the results



Day 5 Task 1

Effects of land use on food webs in grasslands

1. Specify the following model in lavaan
   - For this, if needed, recode the categorical variables in a way appropriate for the analysis
3. Fit the model using data `Food-web_data.csv`
4. Get the fit indices
5. Fill in Standardized Coeficients and R^2 for the model
6. Think about how to interpret the results