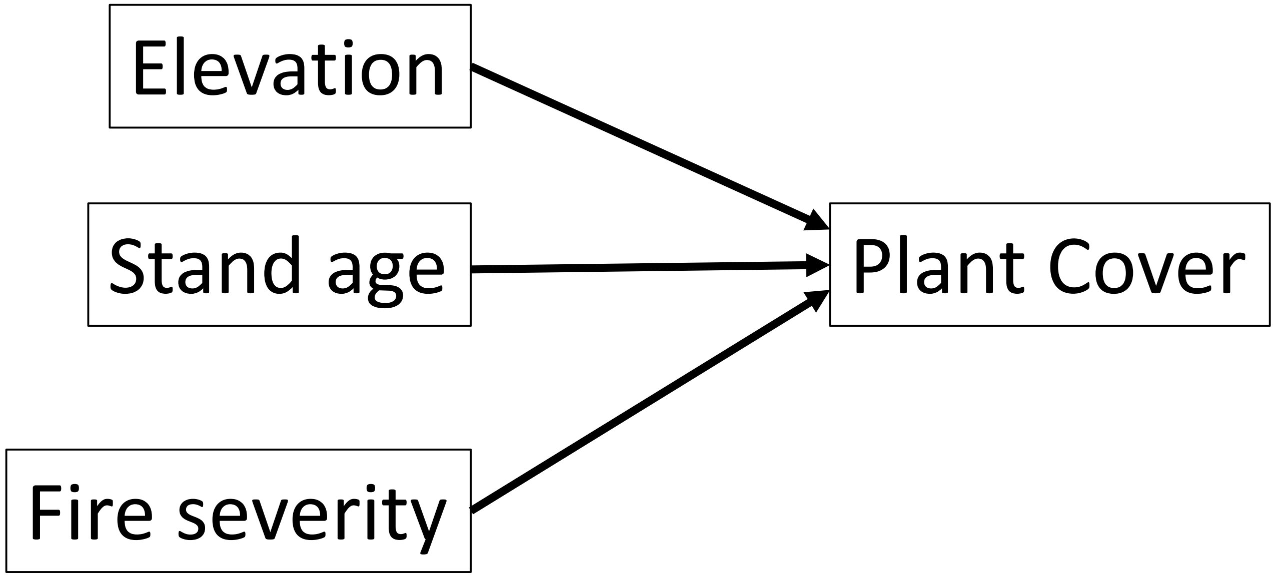
# Structural Equation Modeling

?PX1JHP97587GZ

- Introduction to SEM
- How to interpret pathways
- *Example 1 using lavaan*
- Model fits & saturated models
- *Example 2 using lavaan*
- *Example 3 using piecewise SEM*
- A couple of other things you should be aware of (but I'm not covering)

# Multiple Regression

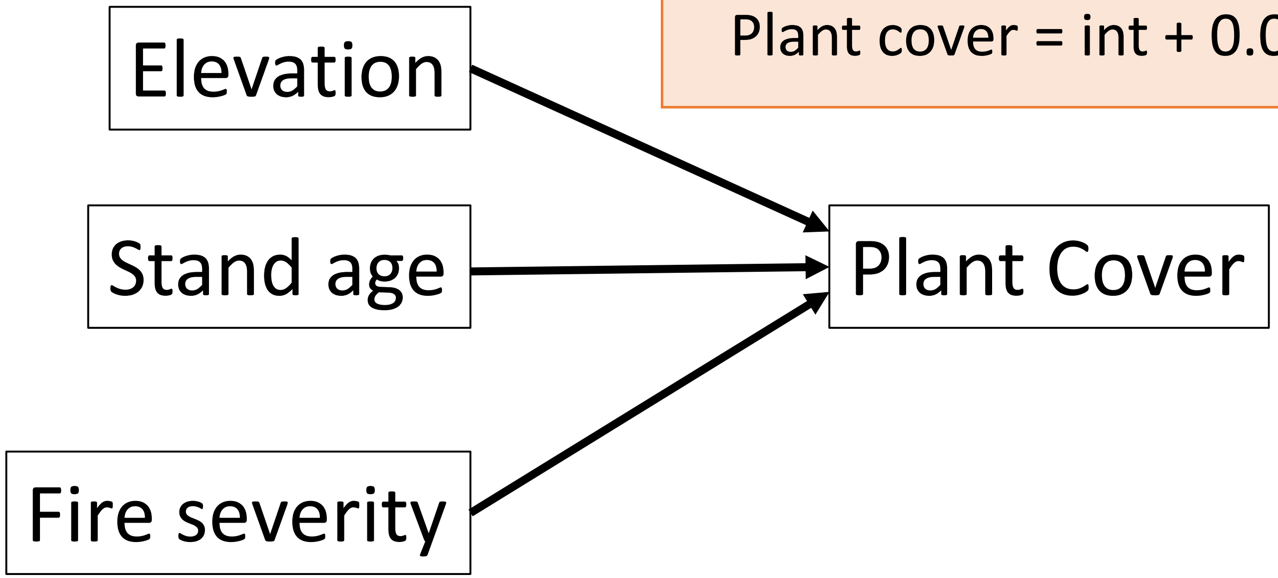
Interpreting results from multiple regression and structural equation models
Grace & Bollen (2005) Bulletin of the Ecological Society of America 86:283-295



# Multiple Regression

```
lm(Plant.Cover ~ Elevation + Stand.age + Fire.severity, data=data)
```

Plant cover = int + 0.038(elevation) + 0.149(age) – 7.96(severity)



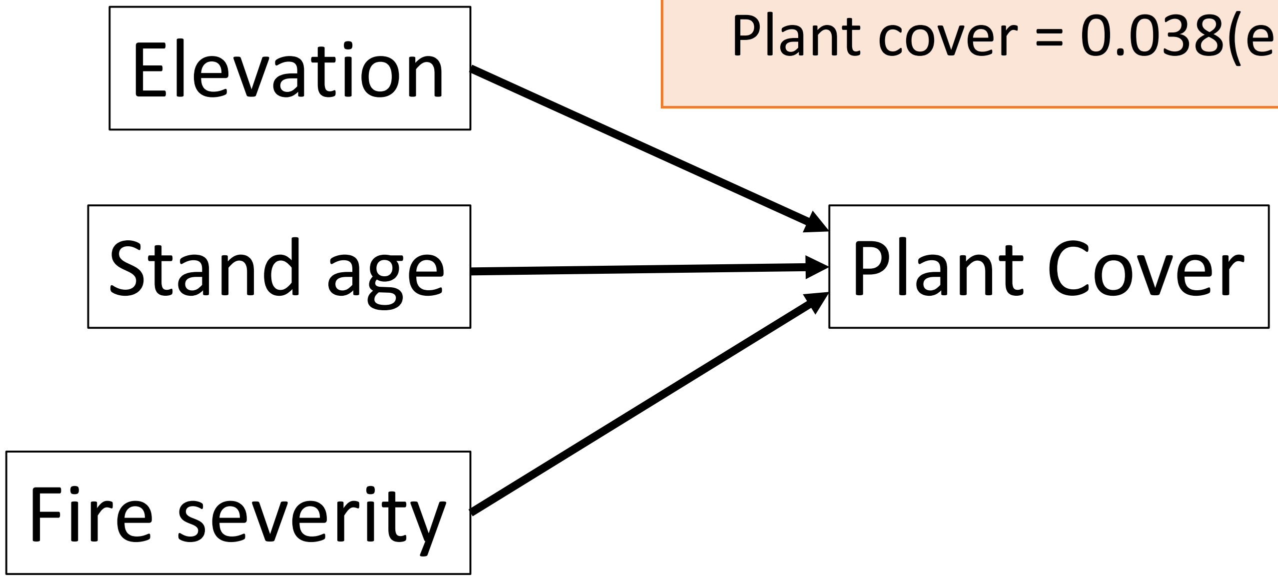

These can not be too correlated!

# Multiple Regression

**Partial regression coefficients**
Effect of elevation on plant cover at average age and fire severity.

```
lm(Plant.Cover ~ Elevation + Stand.age + Fire.severity, data=data)
```

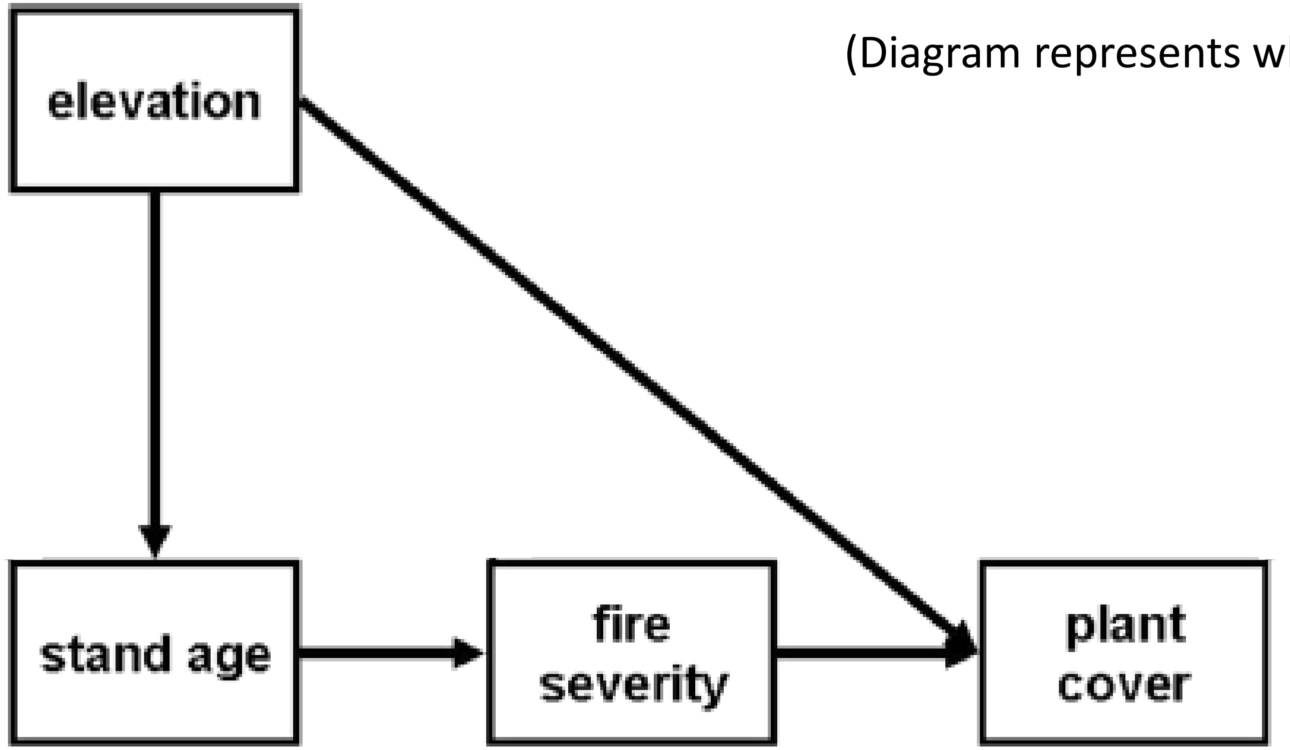
Plant cover = 0.038(elevation) + 0.149(age) – 7.96(severity)




Dictionary definition
"the expected change in the dependent variable associated with a unit change in a given predictor while controlling for the correlated effects of other predictors"

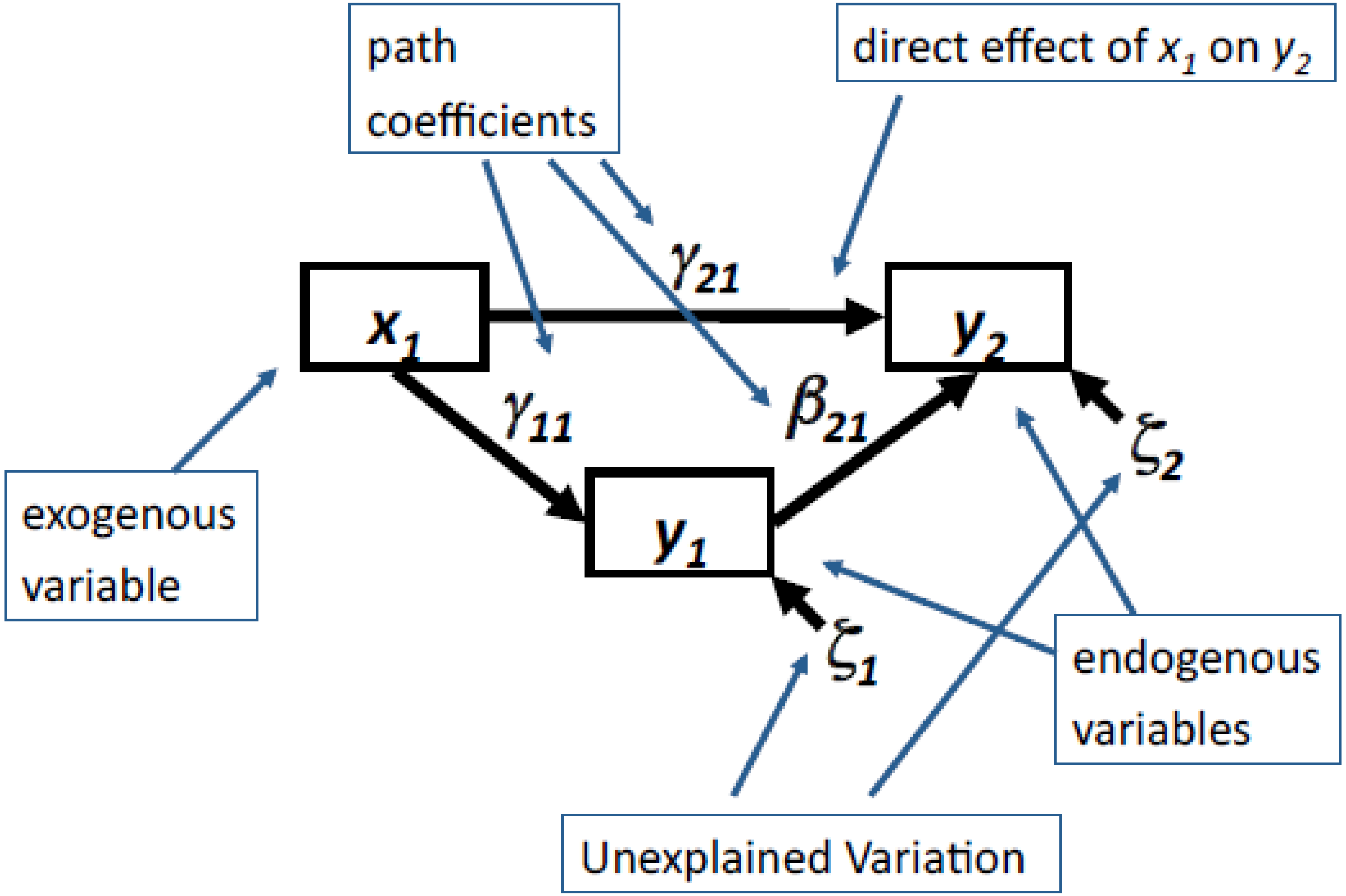
These can not be too correlated!

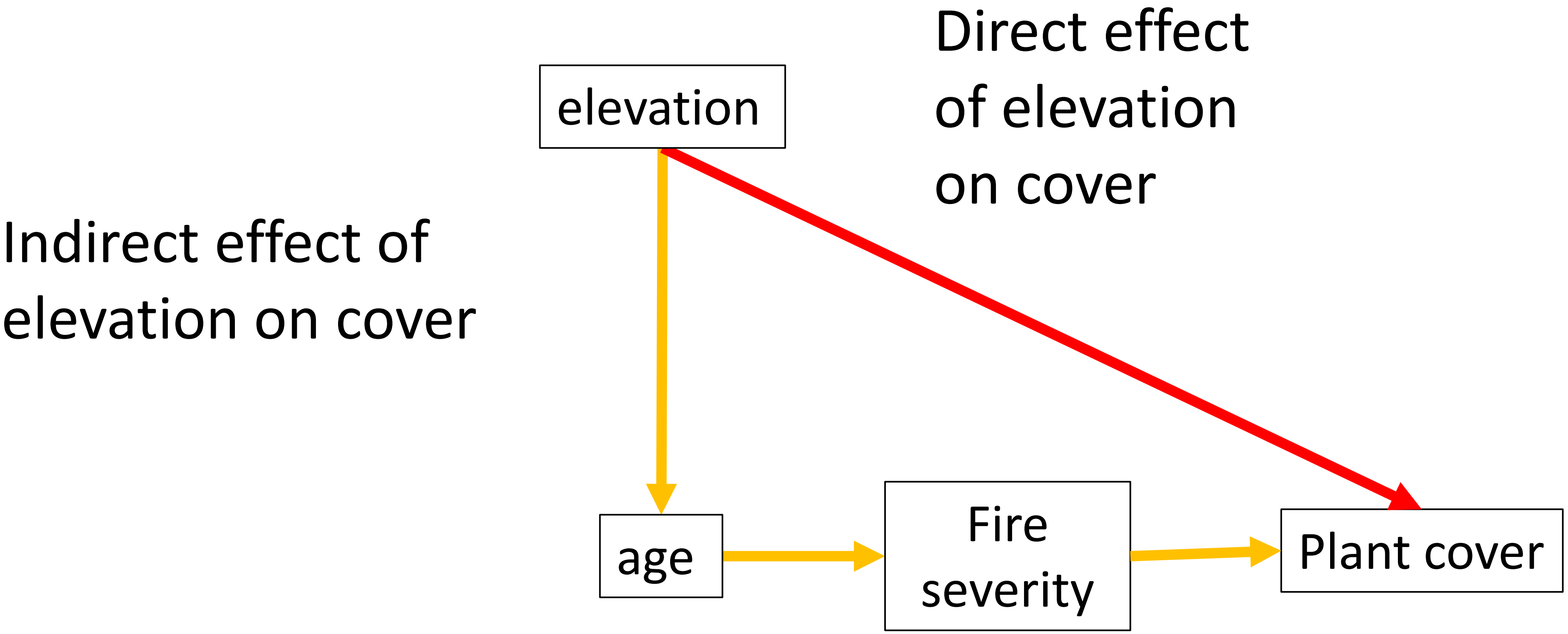
# Better biological hypothesis!




(Diagram represents what a SEM, or path analysis, is!)

# Basic Terminology

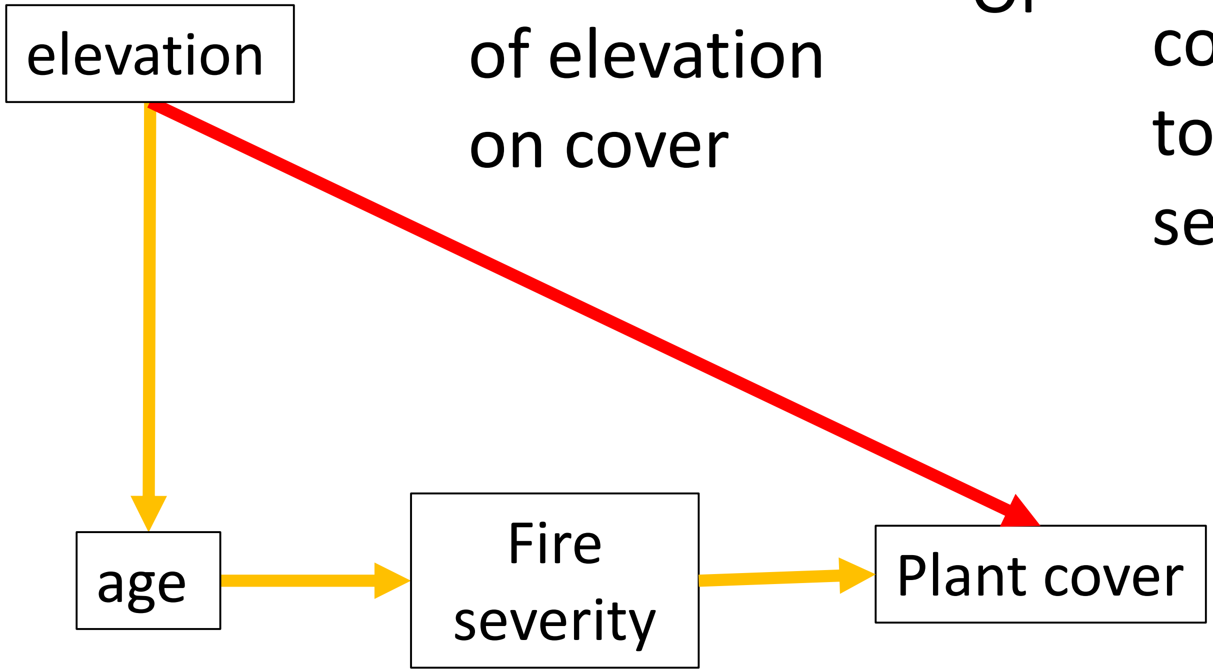



Direct effect of elevation on cover
Indirect effect of elevation on cover
elevation
age
Fire severity
Plant cover

Indirect effect of elevation on cover

Or

Effect of elevation on cover due to age and fire severity

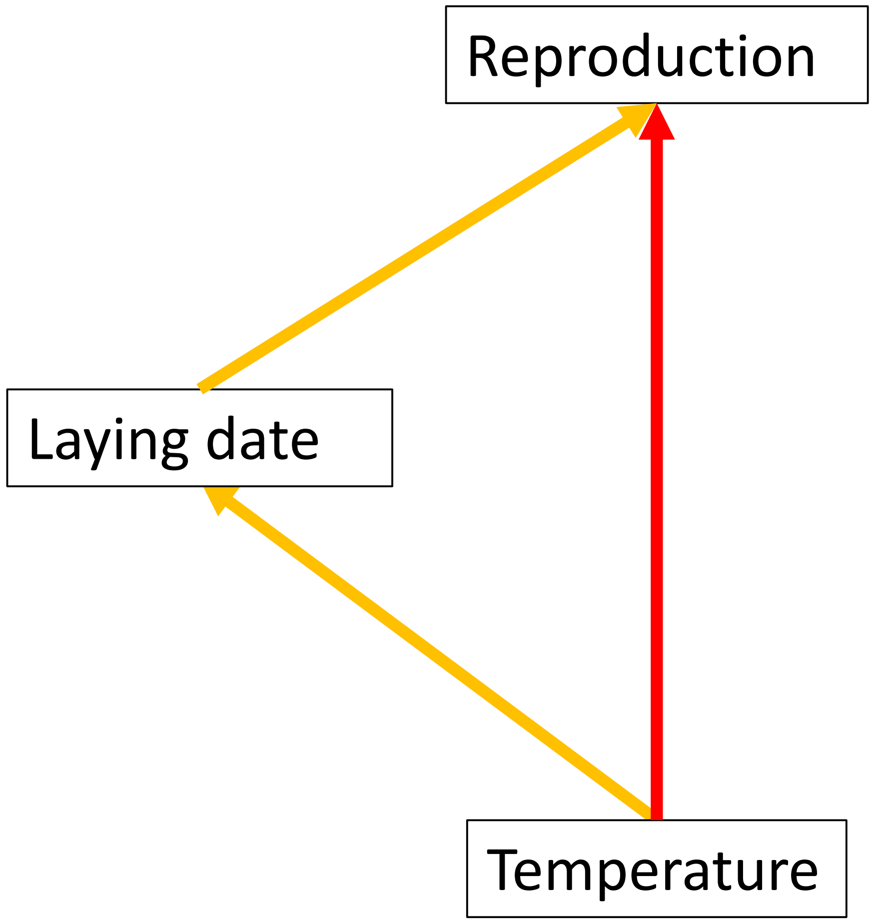


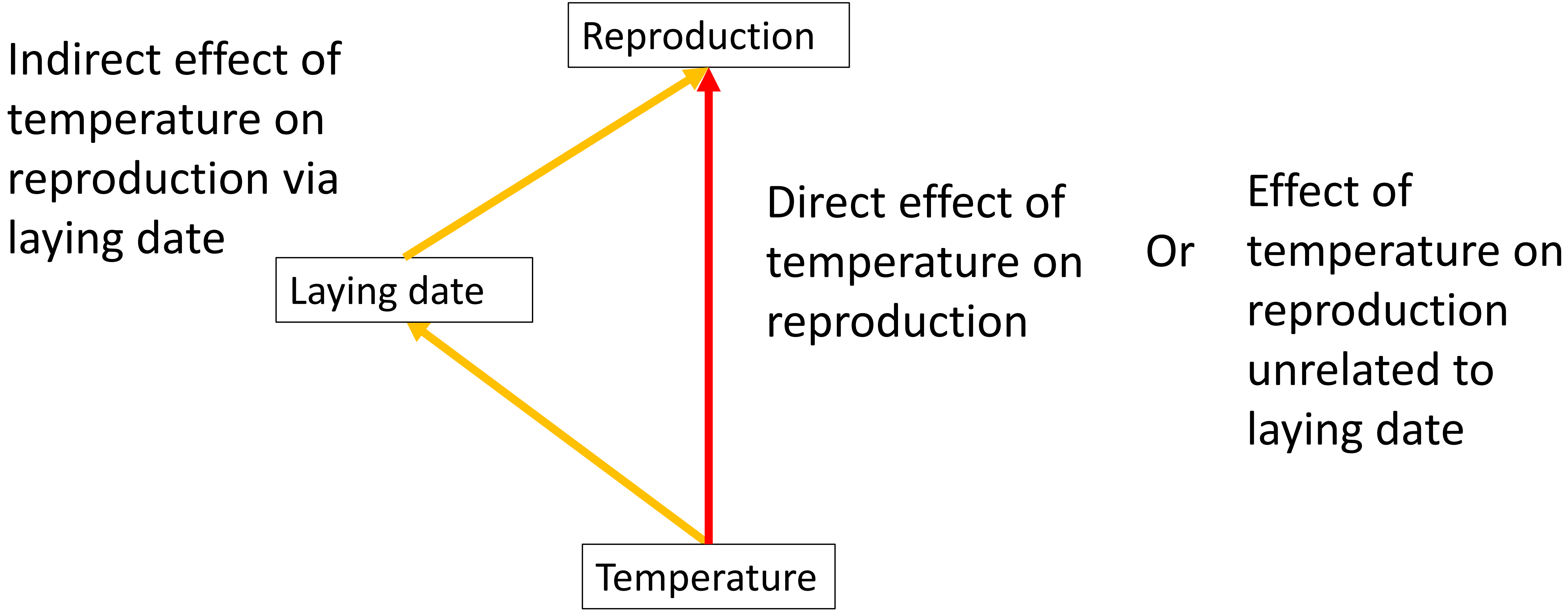

Direct effect of elevation on cover

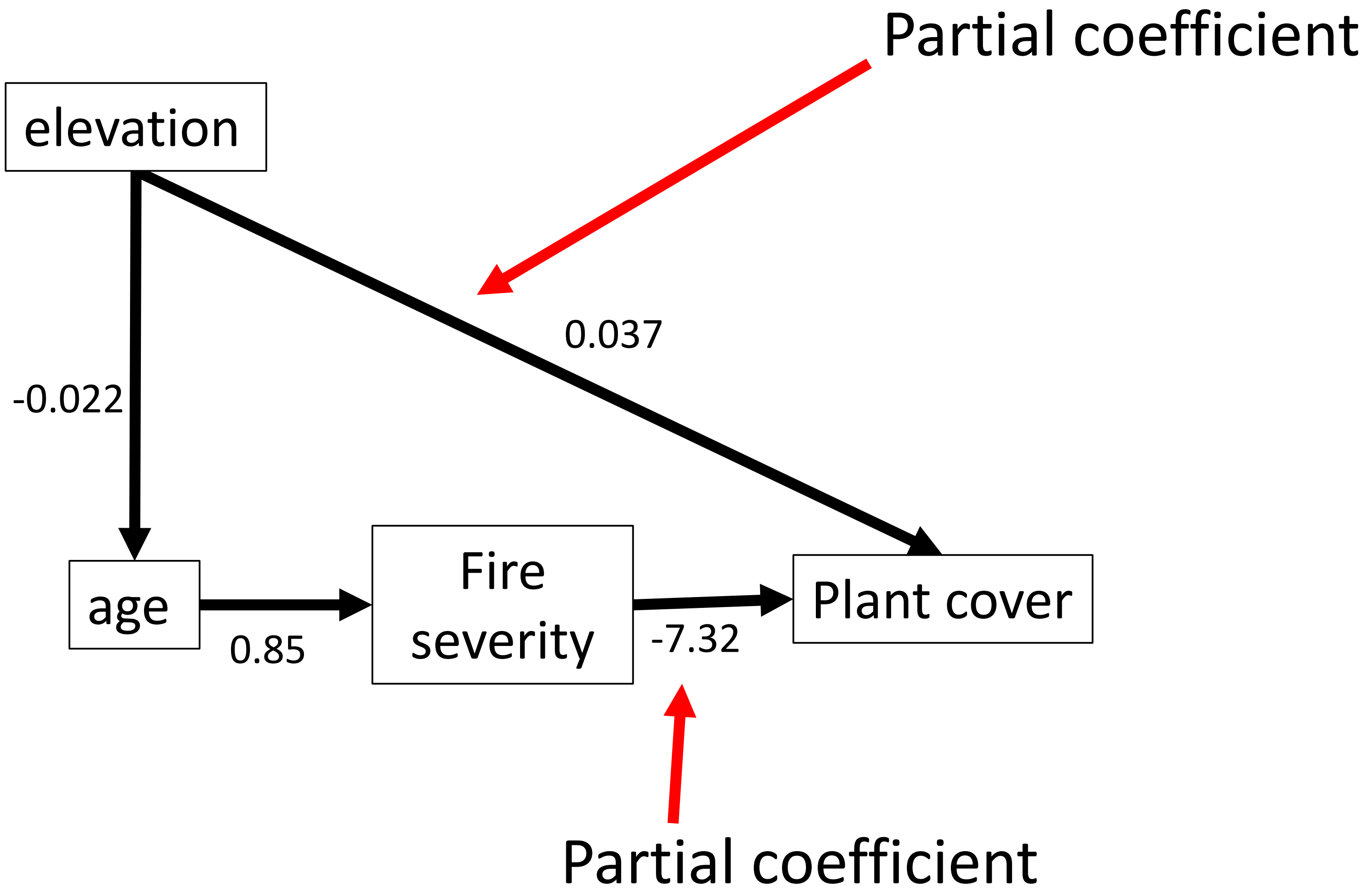
Or

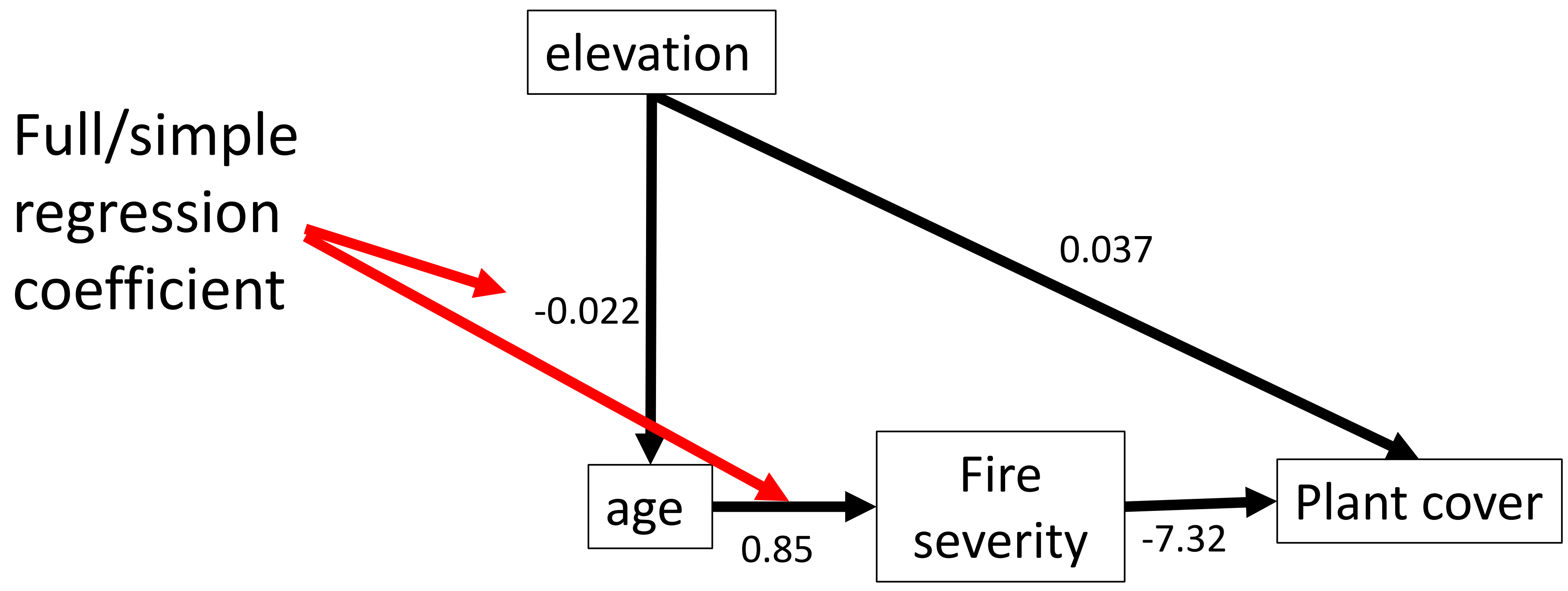
Effect of elevation on cover unrelated to age and fire severity

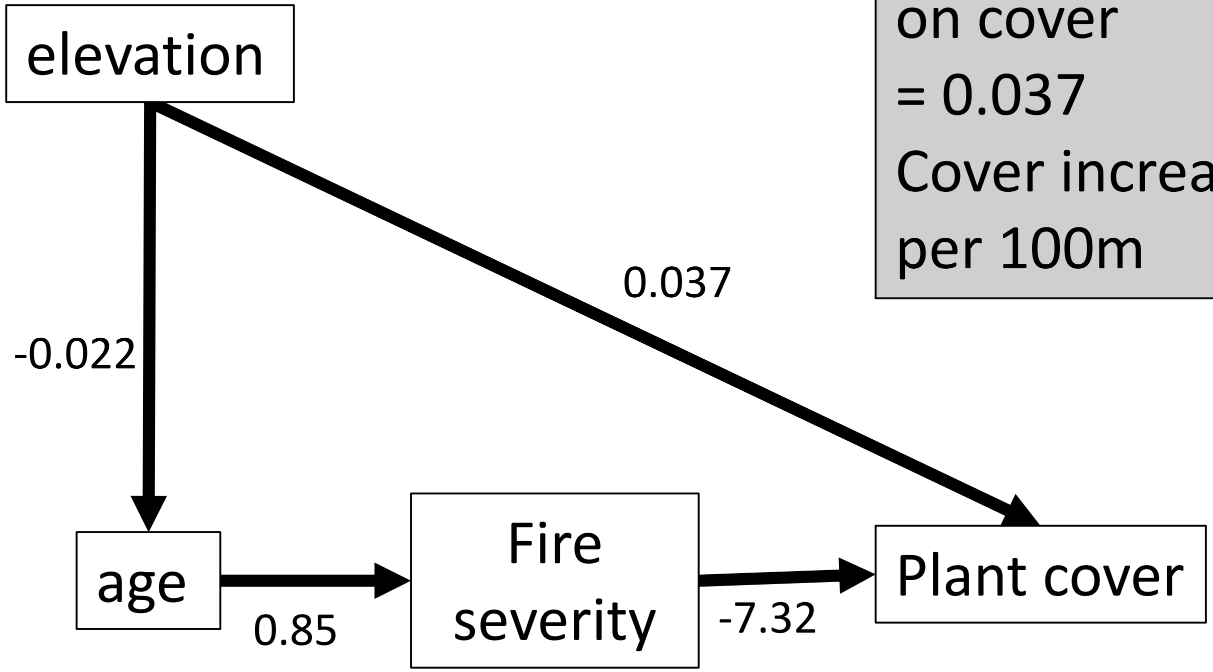
# Another example




Reproduction
Indirect effect of temperature on reproduction via laying date
Laying date
Direct effect of temperature on reproduction
Or
Effect of temperature on reproduction unrelated to laying date
Temperature

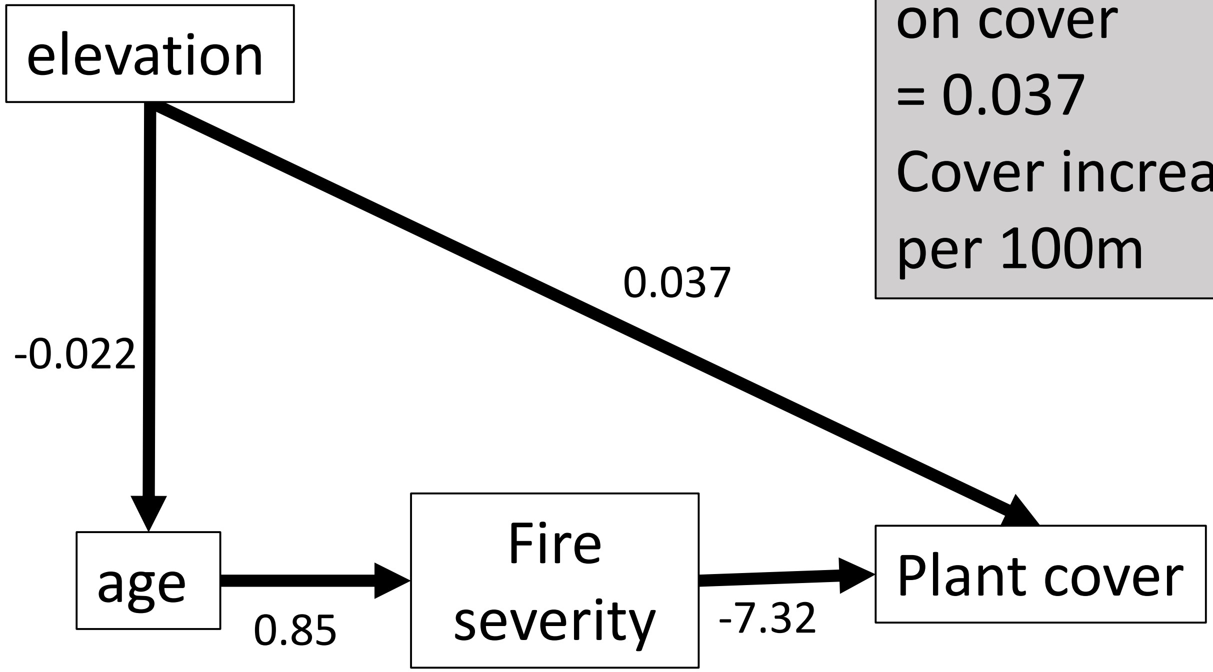

Partial coefficient
elevation
0.037
-0.022
age
Fire severity
Plant cover
0.85
-7.32
Partial coefficient


Full/simple
regression
coefficient
elevation
-0.022
0.037
age
0.85
Fire
severity
-7.32
Plant cover



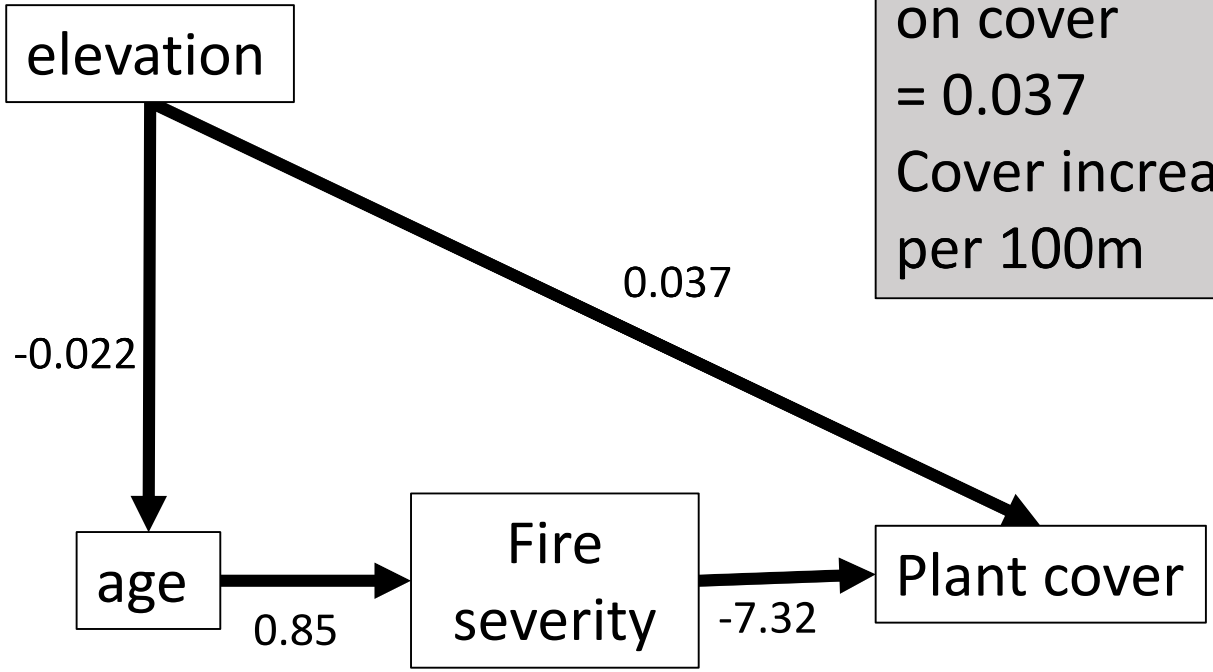

**Direct** effect of elevation on cover
= 0.037
Cover increases by 3.7% per 100m

**Indirect** effect of elevation on cover
= -0.022 * 0.085 * -7.32
= 0.014
Cover increases by 1.4% per 100m




**Direct** effect of elevation on cover
= 0.037
Cover increases by 3.7% per 100m

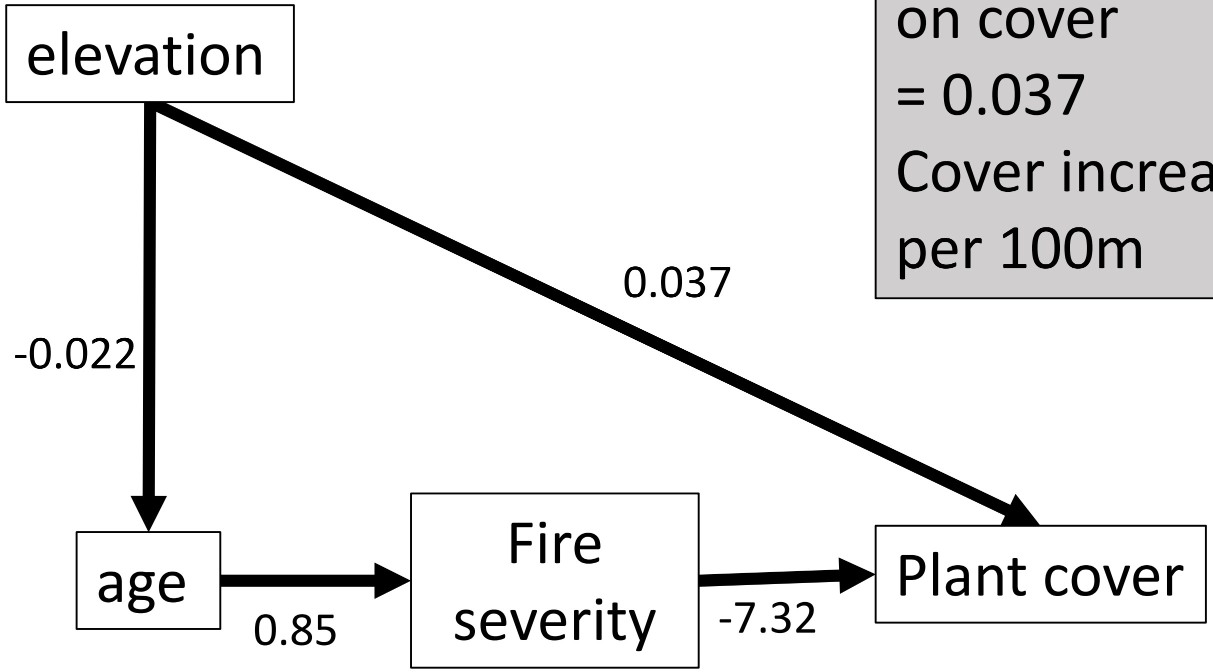
**Indirect** effect of elevation on cover
= -0.022 * 0.085 * -7.32
= 0.014
Cover increases by 1.4% per 100m




**Direct** effect of elevation on cover
= 0.037
Cover increases by 3.7% per 100m

**Total** effect of elevation on cover
= indirect pathway + direct pathway
= (-0.022 * 0.85 * -7.32) + 0.037
= 0.050
Cover increases by 5% per 100m

**Indirect** effect of elevation on cover
= -0.022 * 0.085 * -7.32
= 0.014
Cover increases by 1.4% per 100m

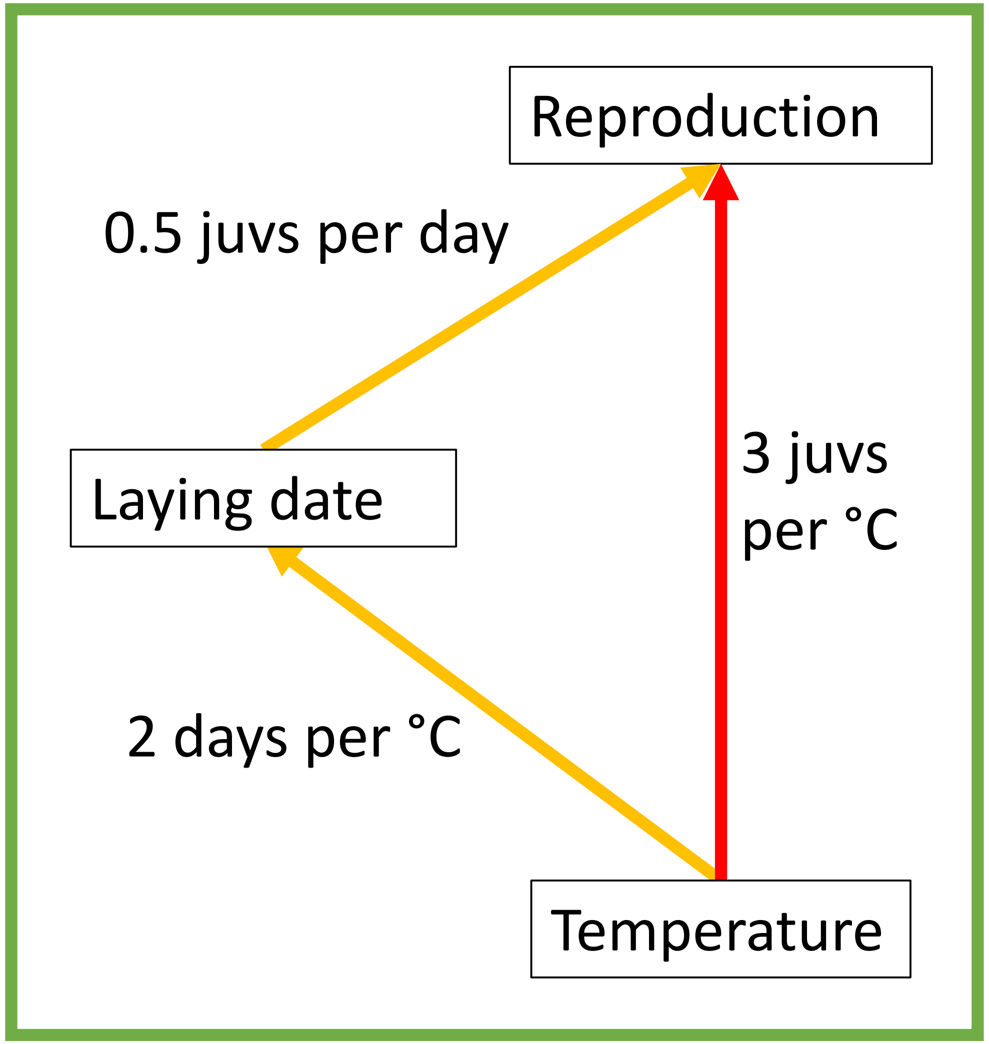



**Direct** effect of elevation on cover
= 0.037
Cover increases by 3.7% per 100m

**Total** effect of elevation on cover
= indirect pathway + direct pathway
= (-0.022 * 0.85 * -7.32) + 0.037
= 0.050
Cover increases by 5% per 100m

If one moved upslope 100m and allowed stand age and severity to vary as it naturally would (i.e. not holding them constant) there would be a net increase in cover of 5%.
Part of this increase (1.4%) would be due to the effects of age and fire severity.

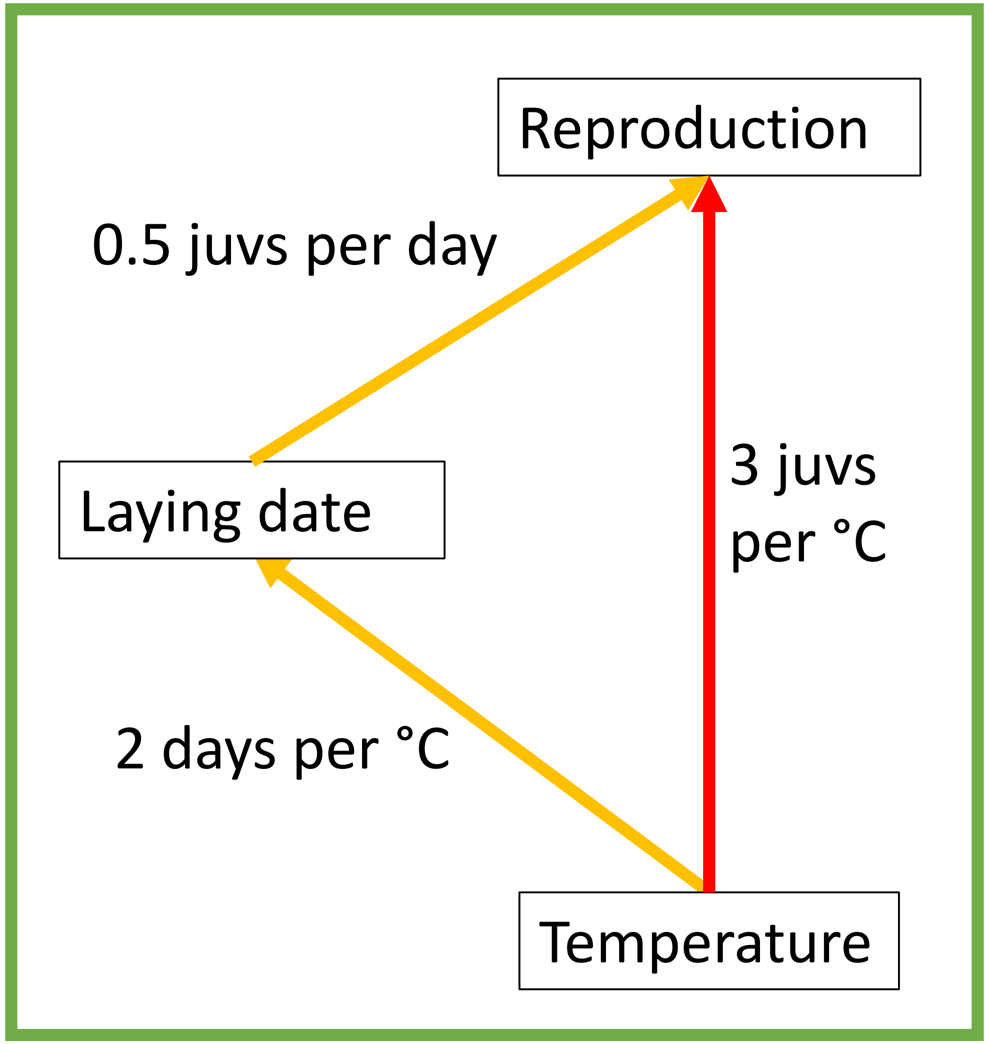
# Perhaps some people think in equations?




$$\frac{dRep}{dTemp} = \overbrace{\frac{dLay}{dTemp} * \frac{\partial Rep}{\partial Lay}}^{\text{Indirect pathway}} + \underbrace{\frac{\partial Rep}{\partial Temp}}_{\text{Direct pathway}}$$

(i.e. Total pathway)

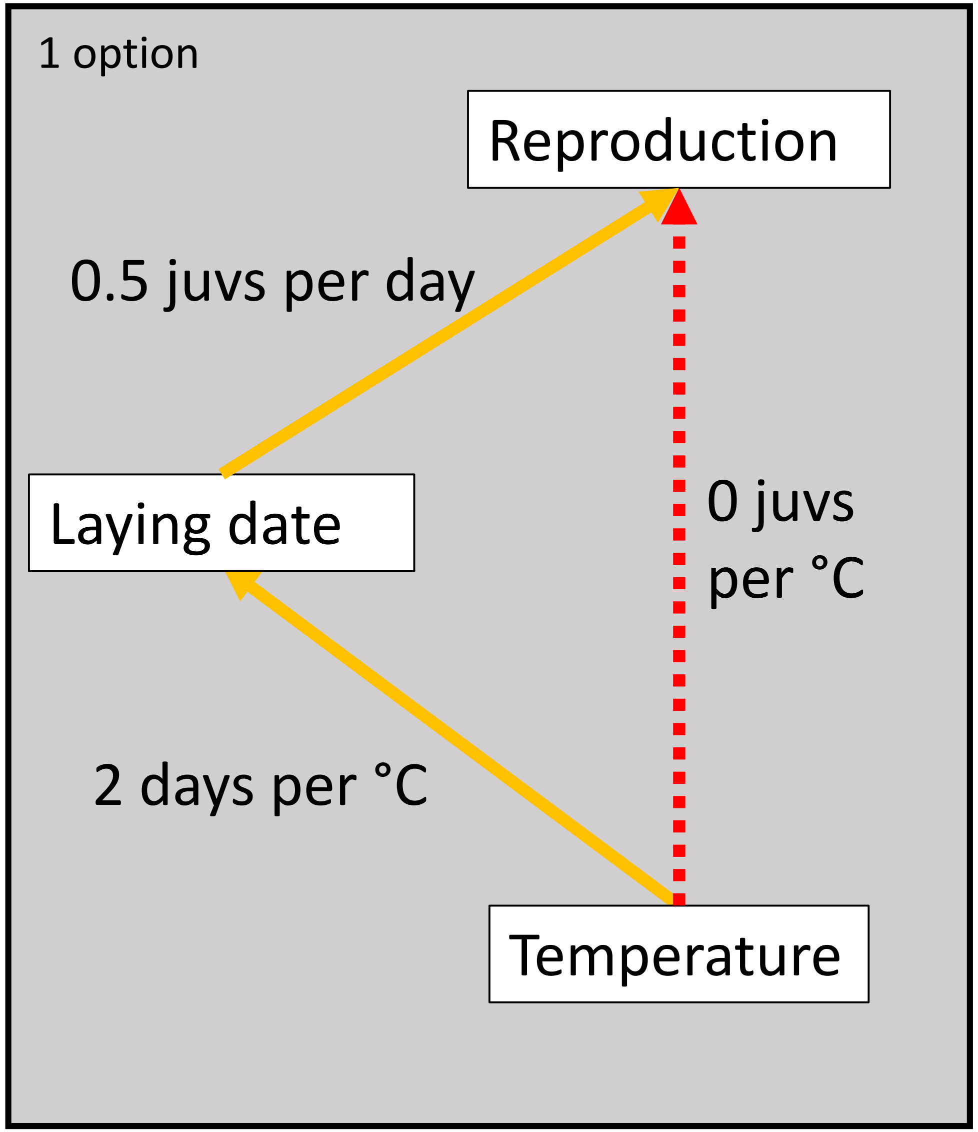
# Quick thought about possible pathways

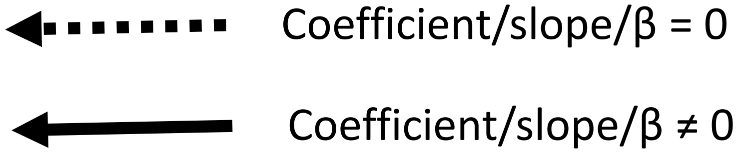



**Questions**

1. What is the indirect effect of temperature on reproduction?
2. What are the units of the indirect effect?
3. What is the total effect of temperature on reproduction?

# Quick thought about possible pathways

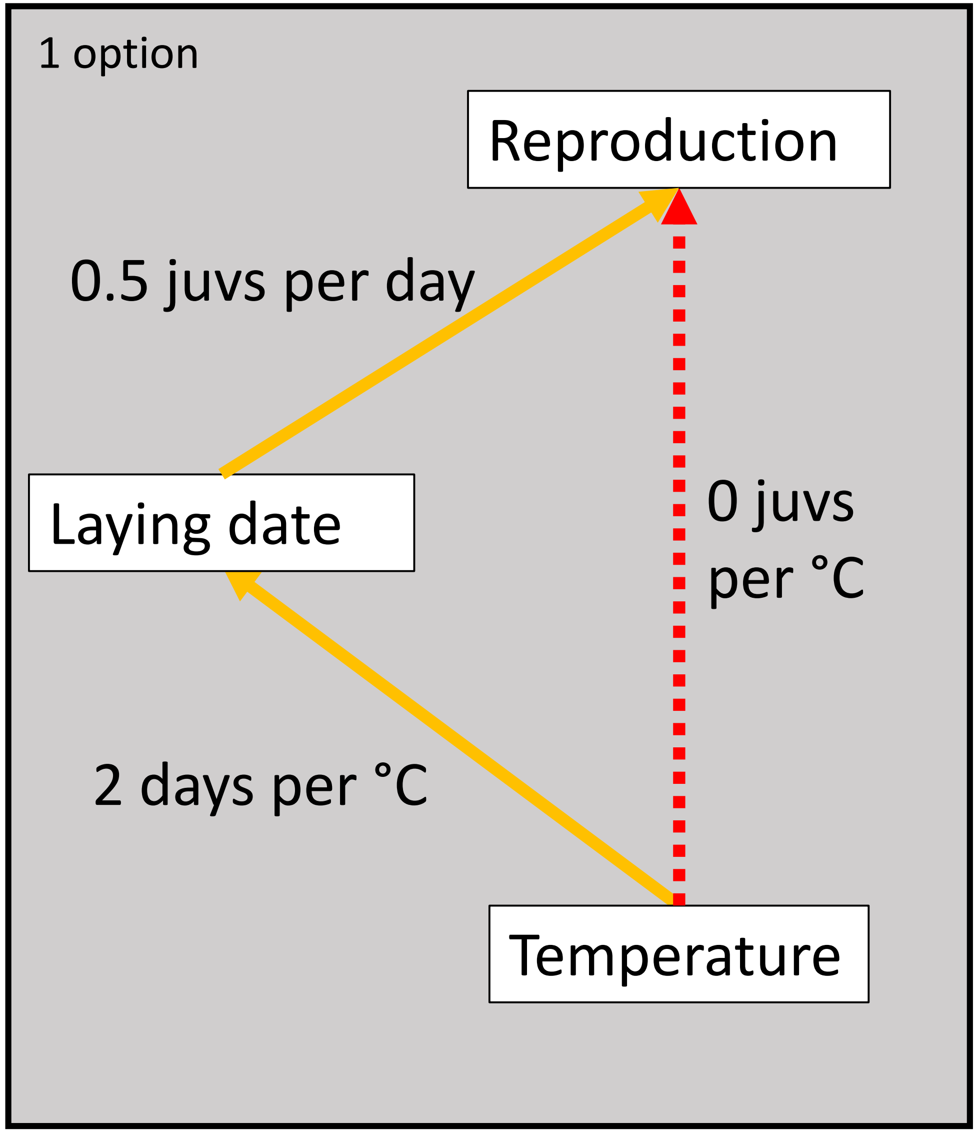


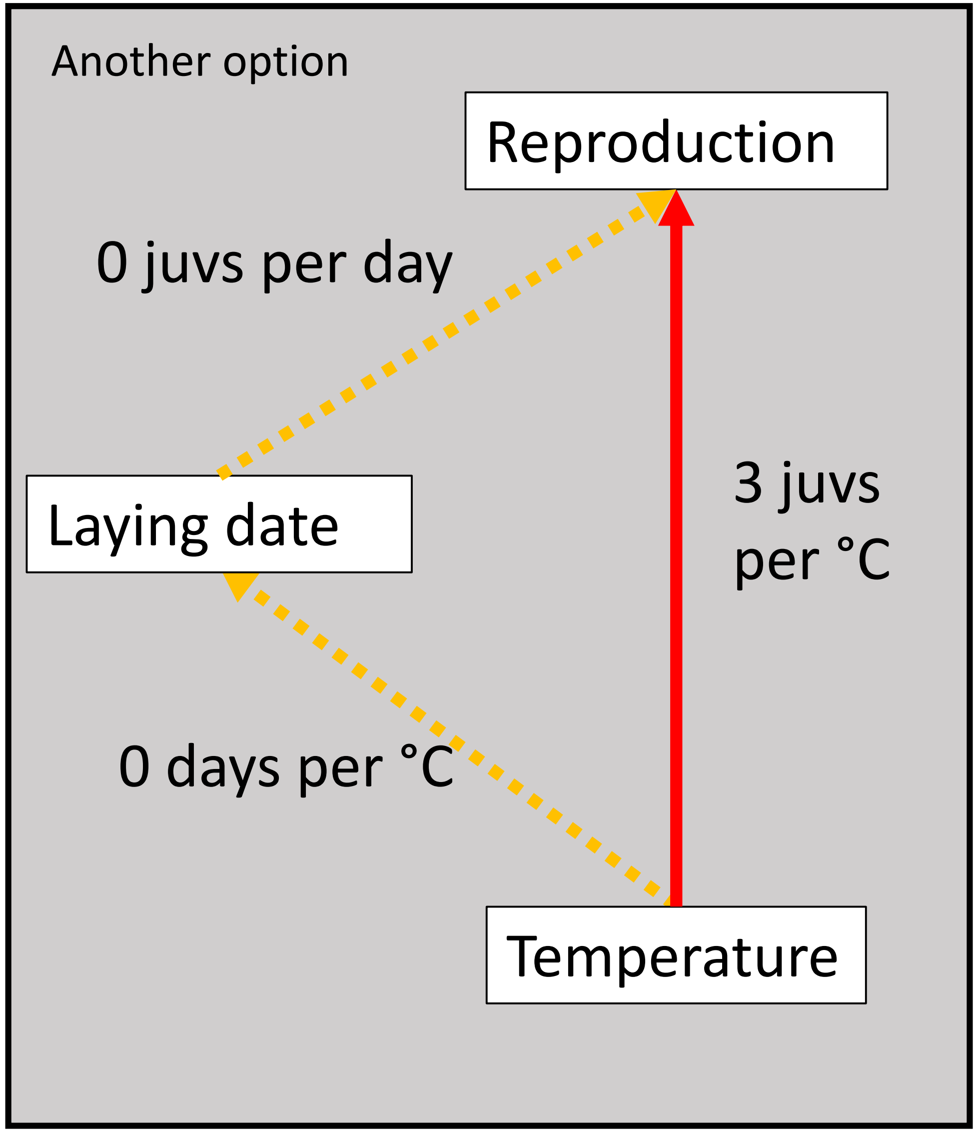




**Questions**

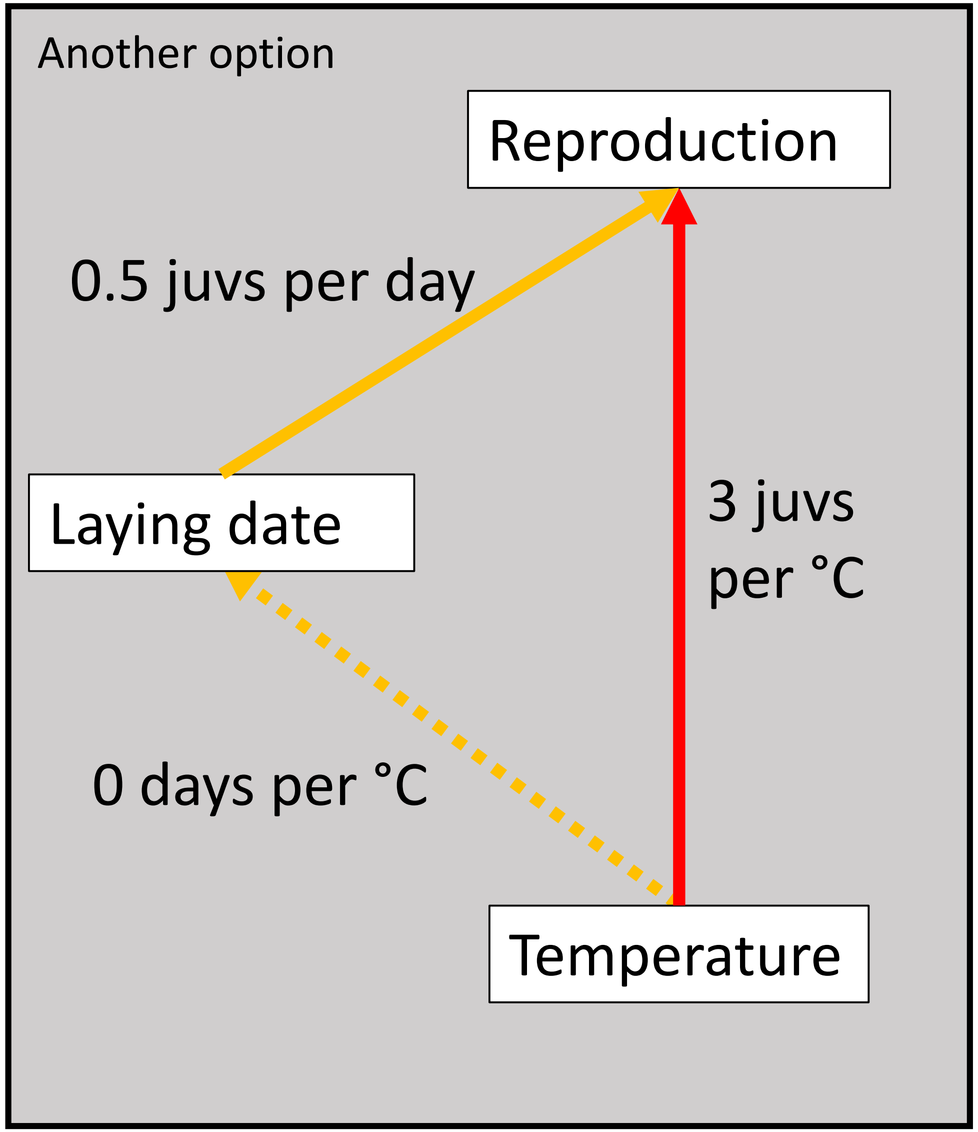
1. What is the indirect effect of temperature on reproduction?
2. What is the total effect of temperature on reproduction?
3. What's your biological interpretation?

# Quick thought about possible pathways

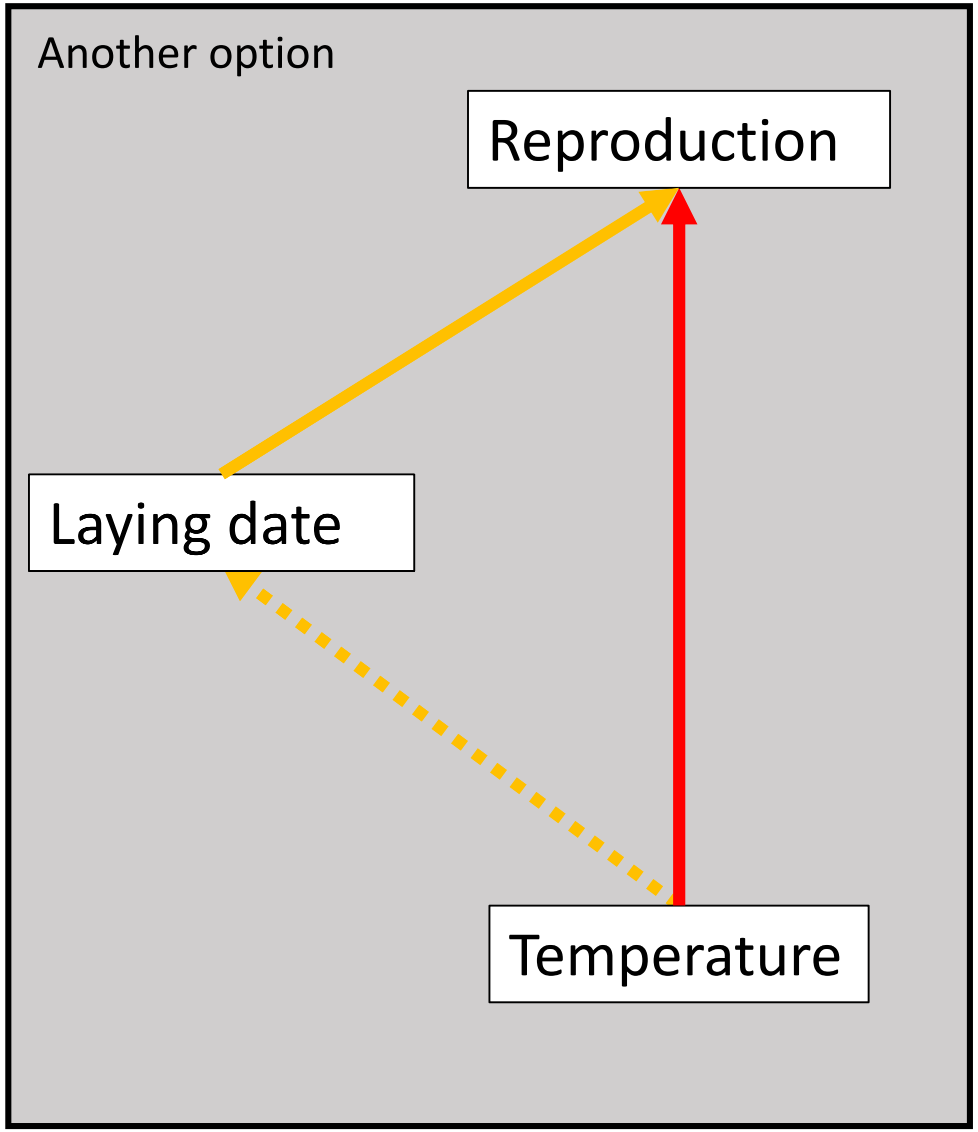


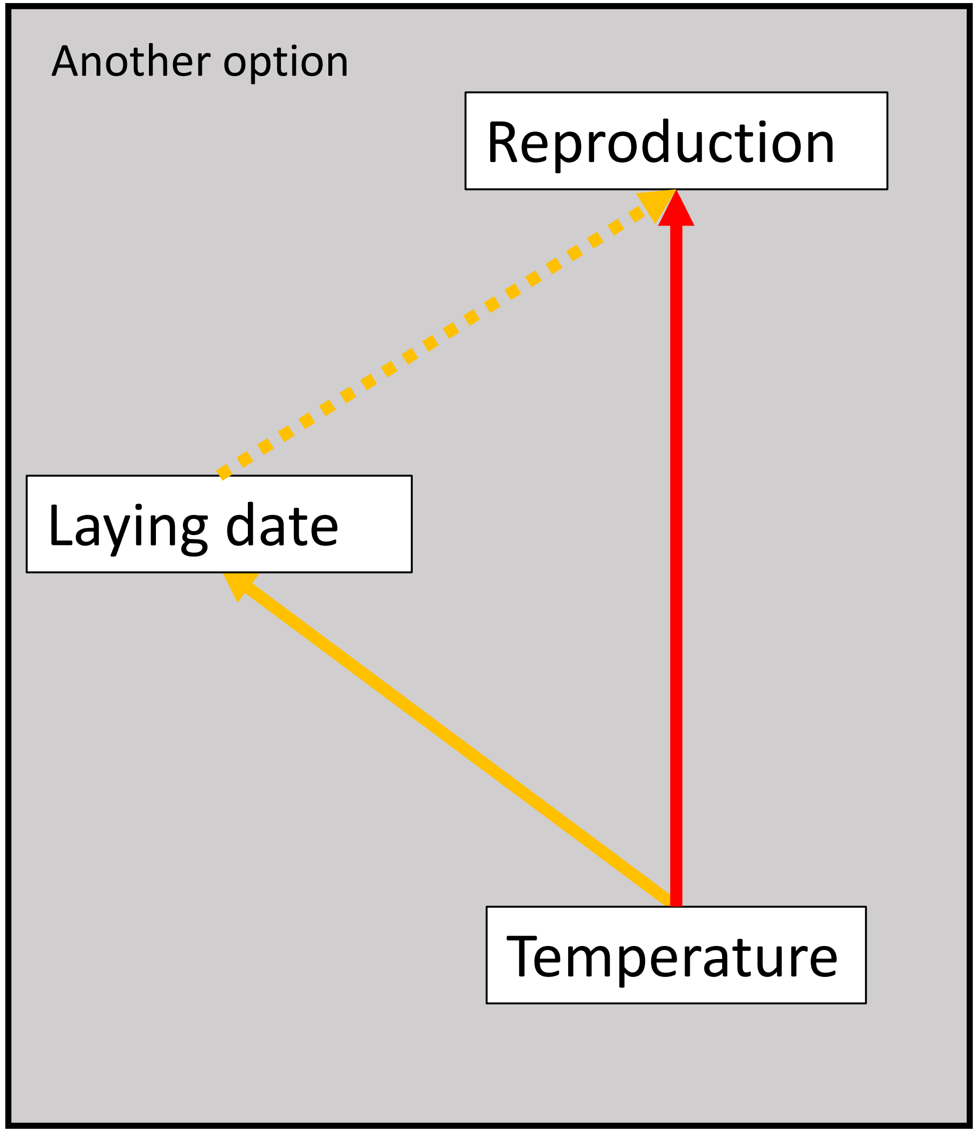



# Quick thought about possible pathways



# Quick thought about possible pathways

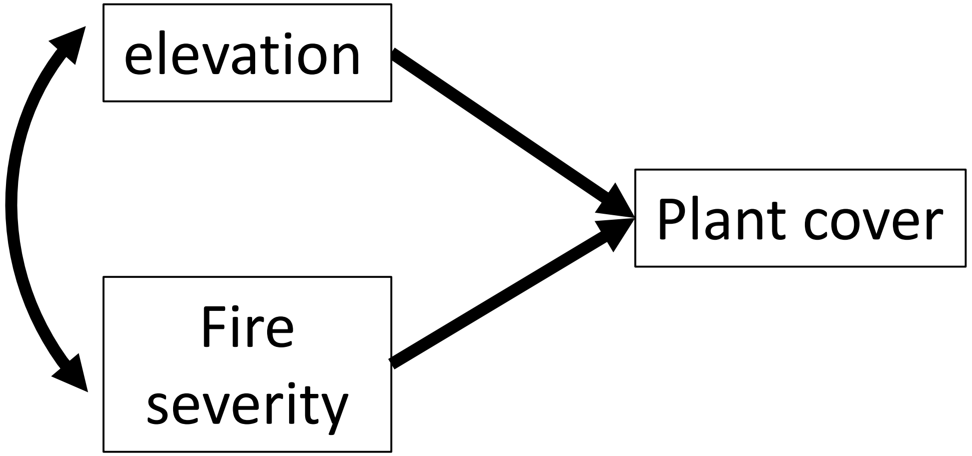


# SEM in lavaan

Borrowed heavily from this great website resource!
http://www.structuralequations.com/LavaanTutorials.html

Let’s try this model first.




NOTE: The path coefficients for unanalysed relationships (curved arrows) between exogenous variables (variables that only predict others) are their correlations (when standardised) or covariances (when unstandardized).

```
> summary(fit, rsq=T, fit.measures=TRUE)
lavaan (0.5-23.1097) converged normally after  16 iterati

  Number of observations                            90

  Estimator                                         ML
  Minimum Function Test Statistic                0.000
  Degrees of freedom                                 0
  Minimum Function Value               0.0000000000000

Model test baseline model:

  Minimum Function Test Statistic               22.383
  Degrees of freedom                                 2
  P-value                                        0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    1.000
  Tucker-Lewis Index (TLI)                       1.000

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)               -529.693
  Loglikelihood unrestricted model (H1)       -529.693

  Number of free parameters                          9
  Akaike (AIC)                                1077.385
  Bayesian (BIC)                              1099.883
  Sample-size adjusted Bayesian (BIC)         1071.479

Root Mean Square Error of Approximation:

  RMSEA                                          0.000
  90 Percent Confidence Interval          0.000  0.000
  P-value RMSEA <= 0.05                             NA
```

```
Standardized Root Mean Square Residual:

  SRMR                                           0.000

Parameter Estimates:

  Information                                 Expected
  Standard Errors                             Standard

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  cover ~
    age              -0.005    0.003   -1.833    0.067
    firesev          -0.067    0.020   -3.353    0.001

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
  age ~~
    firesev           9.319    2.377    3.921    0.000

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)
   .cover             1.122    0.090   12.398    0.000
    age              25.567    1.317   19.410    0.000
    firesev           4.565    0.173   26.356    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .cover             0.078    0.012    6.708    0.000
    age             156.157   23.278    6.708    0.000
    firesev           2.700    0.402    6.708    0.000

R-Square:
                   Estimate
    cover             0.220
```

```
> summary(fit, rsq=T, fit.measures=TRUE)
lavaan (0.5-23.1097) converged normally after  64 iterations

                                                  Used       Total
  Number of observations                            45          48

  Estimator                                         ML
  Minimum Function Test Statistic                0.358
  Degrees of freedom                                 1
  P-value (Chi-square)                           0.550

Model test baseline model:

  Minimum Function Test Statistic               68.824
  Degrees of freedom                                 6
  P-value                                        0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    1.000
  Tucker-Lewis Index (TLI)                       1.061

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)               -115.455
  Loglikelihood unrestricted model (H1)       -115.276

  Number of free parameters                         13
  Akaike (AIC)                                 256.911
  Bayesian (BIC)                               280.397
  Sample-size adjusted Bayesian (BIC)          239.648

Root Mean Square Error of Approximation:

  RMSEA                                          0.000
  90 Percent Confidence Interval          0.000  0.331
  P-value RMSEA <= 0.05                          0.572
```

```
Standardized Root Mean Square Residual:

  SRMR                                           0.011

Parameter Estimates:

  Information                                 Expected
  Standard Errors                             Standard

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  r ~
    fpba    (PpRS)    0.004    0.007    0.599    0.549
    Tmpcntr (Pptm)   -0.001    0.004   -0.280    0.780
  fpba ~
    Laydate (RSly)   -0.027    0.036   -0.748    0.454
    Tmpcntr (RStm)   -0.112    0.196   -0.572    0.567
  Laydate ~
    Tmpcntr (Lytm)   -4.790    0.383  -12.508    0.000

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)
   .r                -0.016    0.037   -0.424    0.672
   .fpba              8.913    4.308    2.069    0.039
   .Laydate         119.344    0.386  309.362    0.000
    Tempcntre         0.068    0.150    0.451    0.652

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .r                 0.001    0.000    4.743    0.000
   .fpba              0.391    0.082    4.743    0.000
   .Laydate           6.667    1.406    4.743    0.000
    Tempcntre         1.010    0.213    4.743    0.000

R-Square:
                   Estimate
    r                 0.009
    fpba              0.013
    Laydate           0.777

Defined Parameters:
                   Estimate  Std.Err  z-value  P(>|z|)
    indPop            0.001    0.001    0.467    0.640
    indRS             0.129    0.173    0.747    0.455
    RStotal           0.017    0.093    0.183    0.855
    Poptotal         -0.001    0.004   -0.262    0.793
```

# Model Fit

If model is a good fit:

- Chi-square value ≥ 0
- P-value (Chi-square): non-significant
- CFI (comparative fit index) ≥  0.95 (some say 0.90)
- RMSEA ≤ 0.06 (excellent), ≤ 0.06 (good), ≤ 0.10 (moderate), >10 (poor fit)
- P-value (RMSEA) likelihood that rmsea is less than 0.05. You want this to be a big number. You want the probability that this number is less than 0.05 to be very high.
- AIC/BIC to compare among models (smaller is better)

# Model Fit

Poor model fit suggests that there is substantial covariation between exogenous variables that isn't explained by your model.

The reason for this might be that there additional latent factors, regression terms, covariance terms etc. that your model is missing, and/or parameters it currently has which are incorrectly specified.

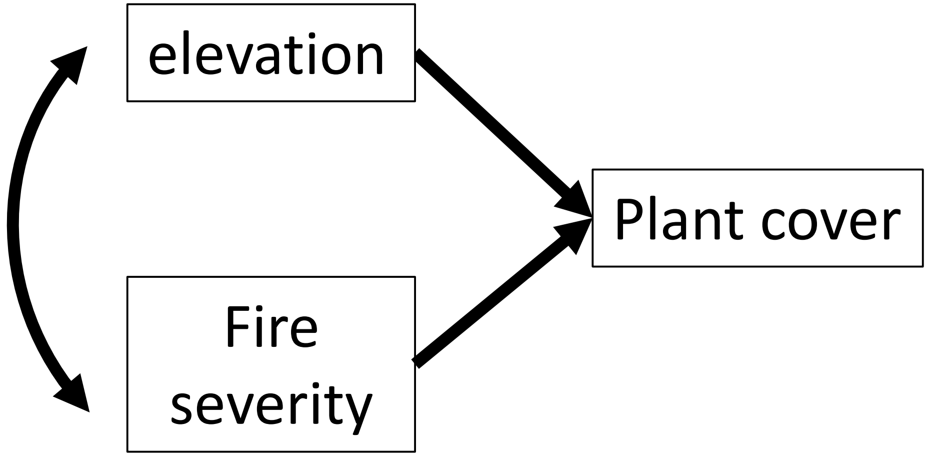
Presumably, if the "correct" model was specified, the parameter estimates would be of different magnitudes (even for those parameters included in both your model and the "correct" model).

# Saturated/just-identified models

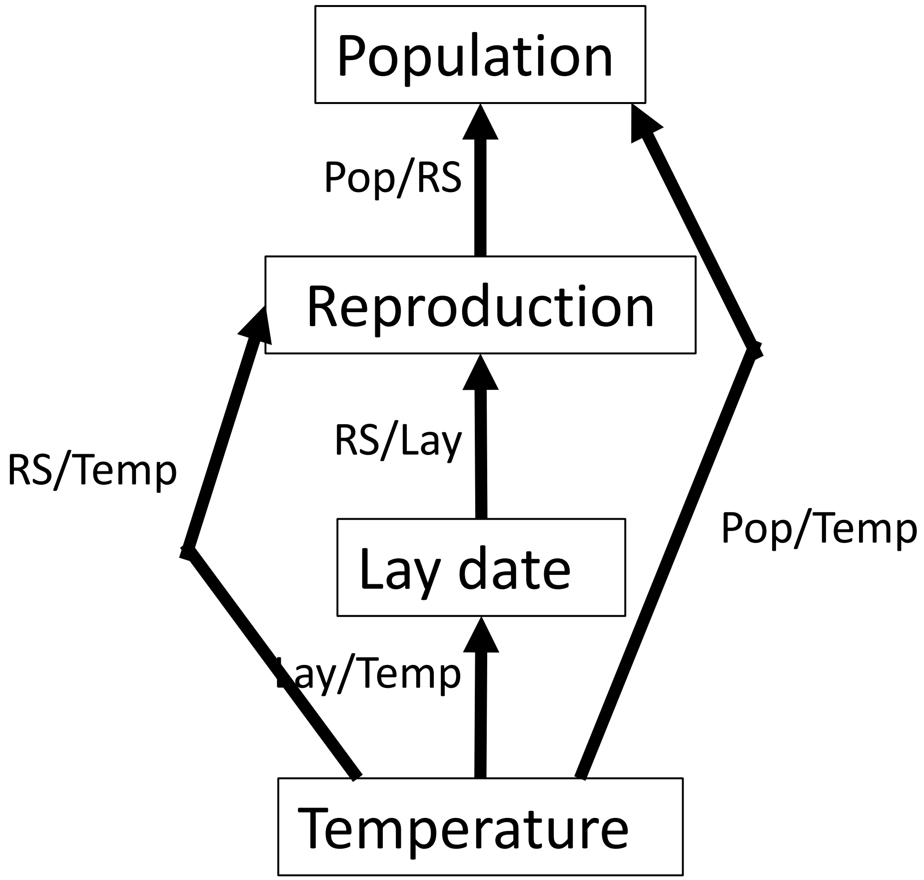
A *just-identified* model is a model that utilizes all of the uniquely estimable parameters. This type of model will always result in a “perfect fit” to the empirical data. Since there is no way one can really test or confirm the plausibility of a just-identified model (also referred to as a saturated model), this type of model is also problematic.

- The number of parameters estimated by the model equals the number of variances and covariances in the data matrix.

```
lavaan (0.5-23.1097) converged normally after  13 iterations

  Number of observations                            90

  Estimator                                         ML
  Minimum Function Test Statistic                0.000
  Degrees of freedom                                 0

Parameter Estimates:

  Information                                 Expected
  Standard Errors                             Standard
```



# Bigger model in lavaan




Write out the equations!

Another way to think about this:

$$\frac{dPop}{dTemp} = \frac{dLay}{dTemp} * \frac{\partial RS}{\partial Lay} * \frac{\partial Pop}{\partial RS} + \frac{\partial RS}{\partial Temp} * \frac{\partial Pop}{\partial RS} + \frac{\partial Pop}{\partial Temp}$$

# Complex data

- Non-normal data (poisson, logistic etc.)
- Need to account for non-independent measurements?
- Want random intercepts, slopes?

Package piecewiseSEM in R is useful for this.

Download from github

```
library(devtools)
install_github("jslefche/piecewiseSEM@2.0")
library(piecewiseSEM)
```

Still a package under progress

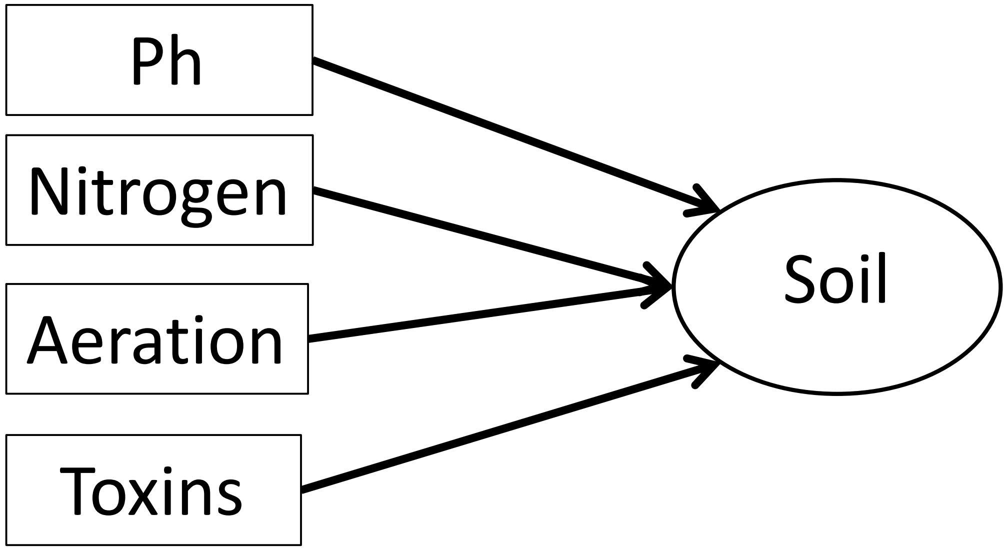
# Things you should know!

(that I’m not covering)

# Latent variables

Latent variables are hypothetical or theoretical variables (constructs) that cannot be observed directly. Latent variables are of major importance to most disciplines but generally lack an explicit or precise way of measuring their existence or influence.

e.g. personality type



# Piecewise SEM

Doesn’t solve equations simultaneously, therefore you need to account for error – don’t want propagation of error.

- Bootstrapping technique to solve this issue! Just ask me for the code.
- Currently being worked into the package itself, but not quite there yet.

Can’t calculate combined paths if one pathway is non-linear!