Linear regression models

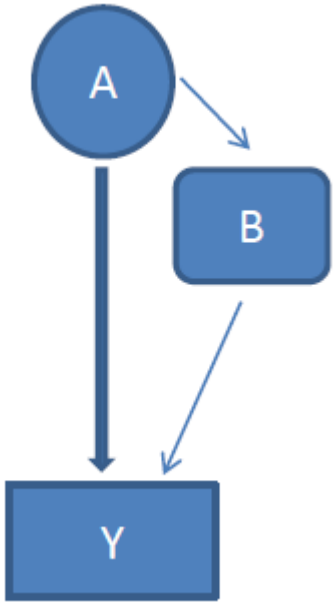
Modelling in physical geography, 5cr, 30.10.-30.11.2017

Essentially, all models are wrong, but some are usefull.

- Box, 1986

Modelling in physical geography

- Physical geography aims to *describe, explain and predict* nature's complex spatial patterns
- We seek to find *general trends* from *empirical* data that helps to understand environmental processes
- **Model = a simplification of reality**
- Statistical model = simplified, mathematically-formalized way to approximate reality (reality in this case, is our data)
- The concept and inference of *causality*, identification of links between variables that are supported by the theory
- One of the most common statistical modelling technique is *linear regression*



Modelling in physical geography

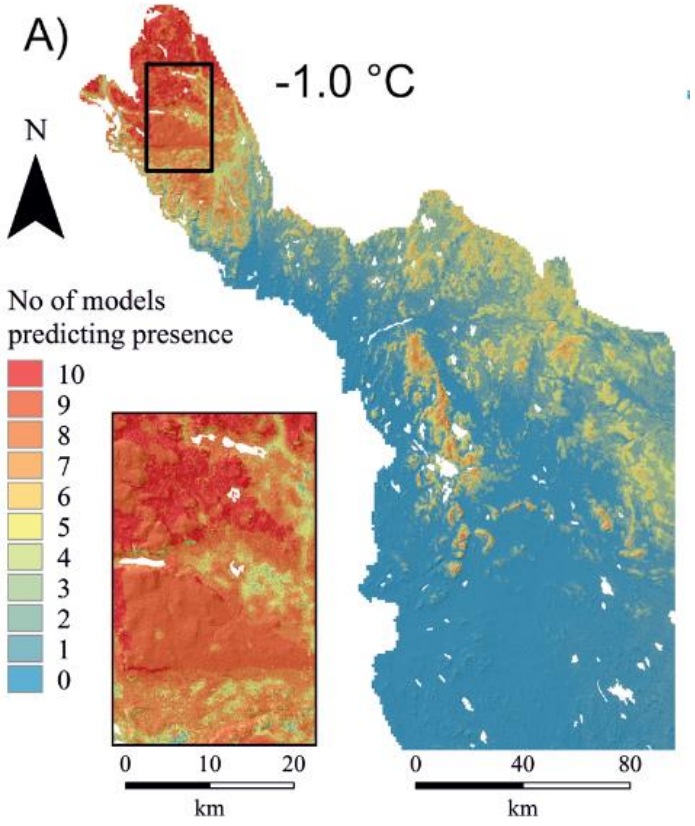
- Modelling can be used to test existing theories and hypothesis
- *Conceptual models* structure research questions and aims, and links between variables
- When modelling, one aims to maximize either:

  - ”realism” (model is built on key causal relationships)

  - accuracy (small errors between modelled and observed values)

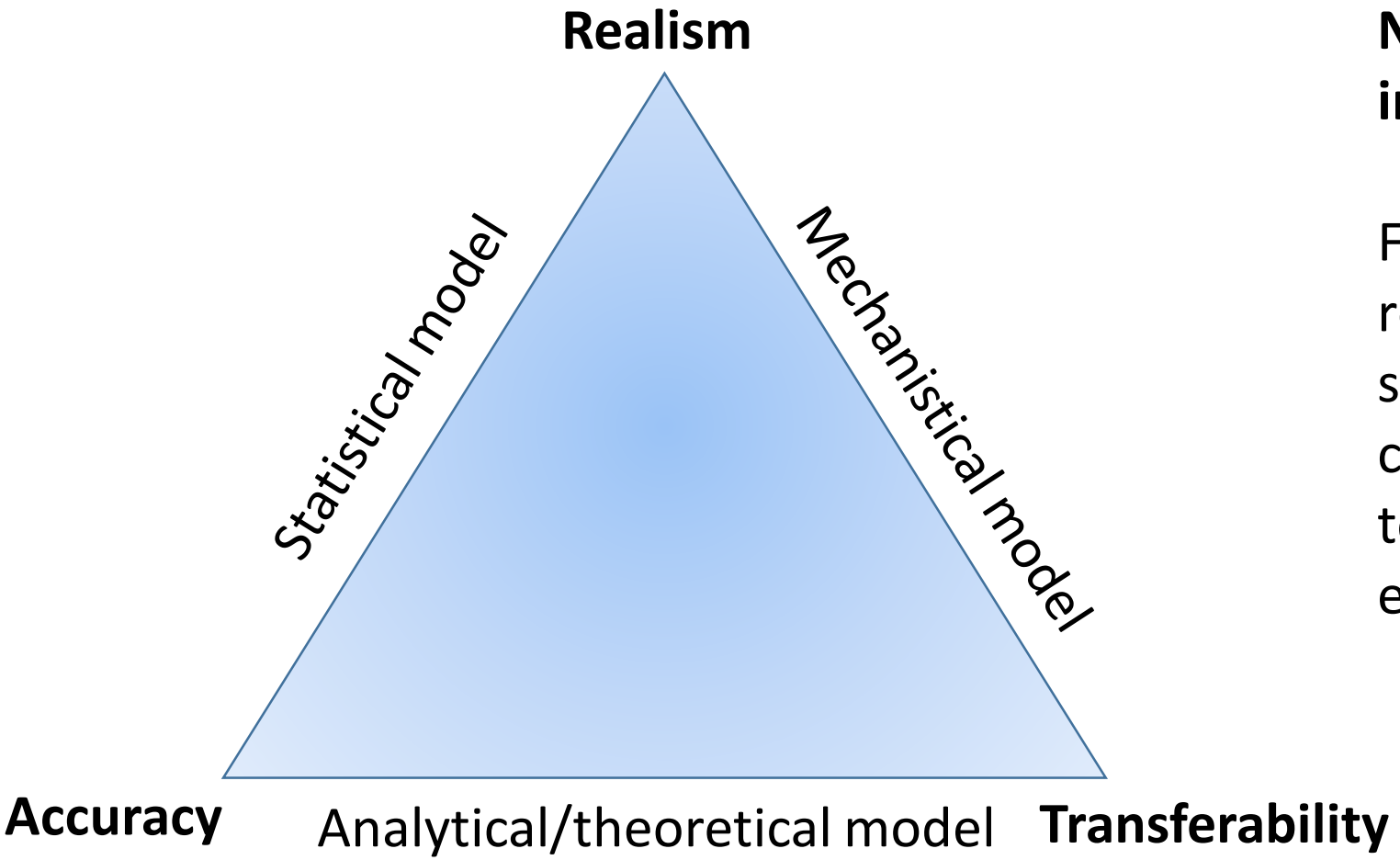
  - transferability/generality (how well the model works in another environment?)

- **Models are always imperfect presentation of reality**
- Uncertainty derives from empirical data (e.g. sample size), number of variables, modelling algorithm, **interpretation**, ...



Number of different modelling methods to predict presence of cryoturbation

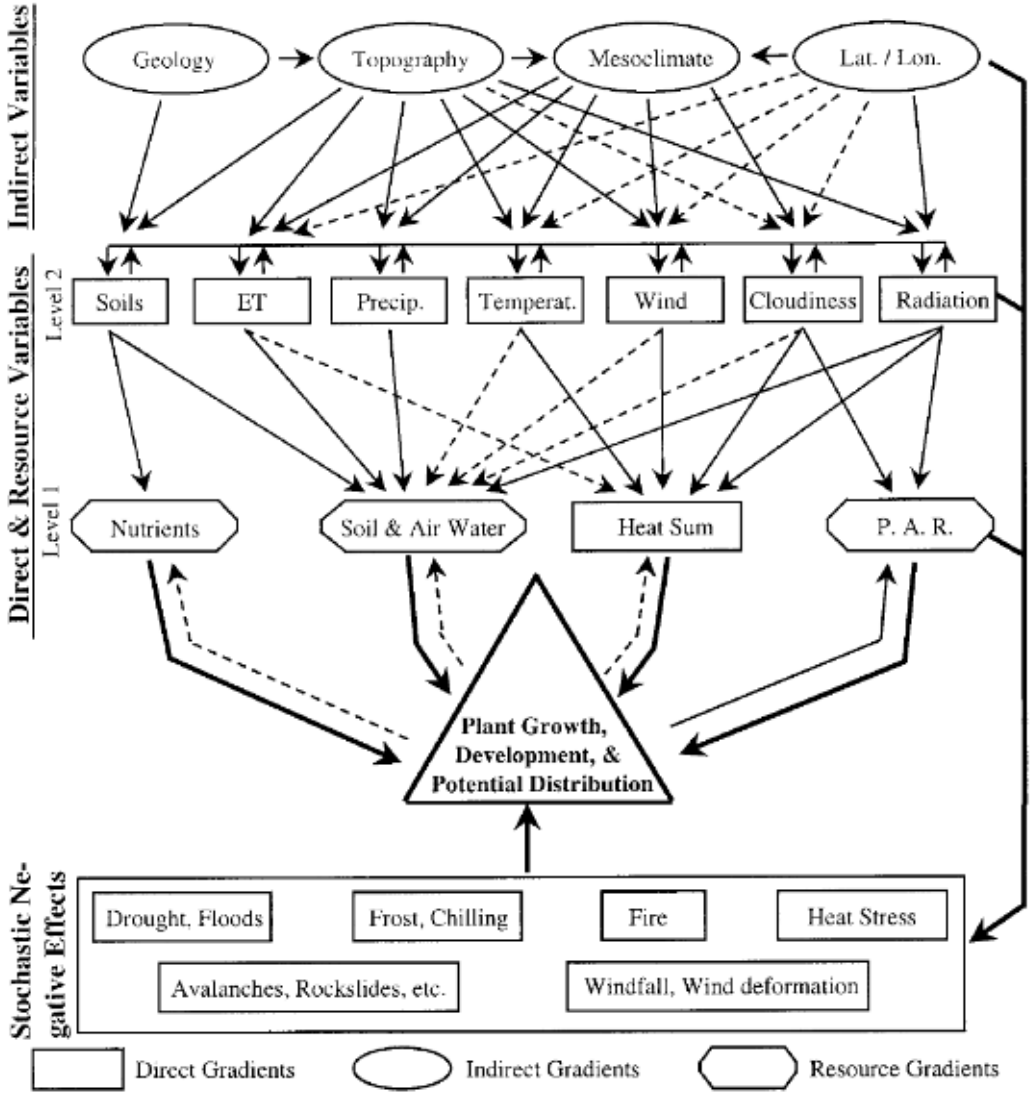
Holy trinity in modelling



No model is optimal in every sense!

For example, a realistic model can be so complex that it can't be generalized to another environments

A complex model – modelling plant distributions



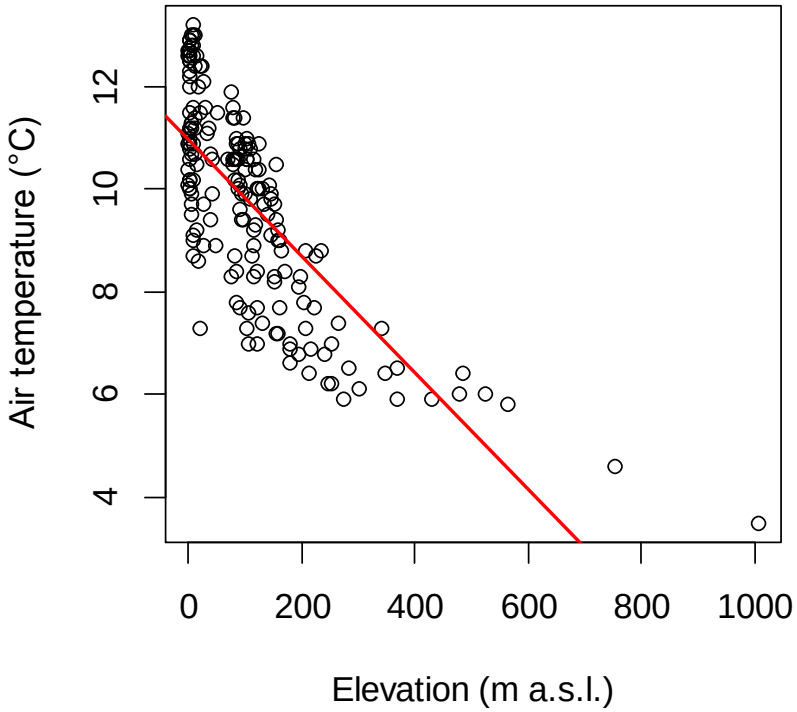
Indirect factors, such as topography and location, control direct factors and resources

Resources and direct factors, such as growing degree days, moisture and nutrients, are of key importance for species

Stochastic factors, such as wind, frost, storms, are difficult to forecast and account for in a model

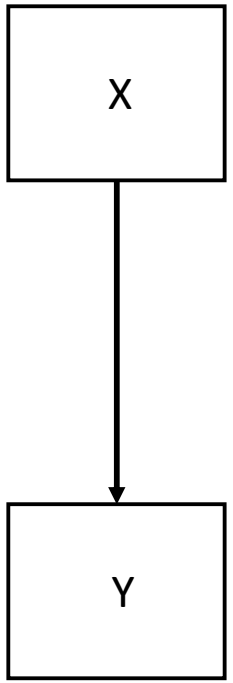
Linear regression - rationale

- The aim of linear regression analysis is to describe and measure *the average relationship* between response **y** and predictors (**x1**, ..., **xn**)
- In practice, a line is *fitted* to the observation data that describes as well as possible the relationship between the two variables
- The model can be used to predict response variable, when predictors' values are known



Linear regression – key terminology

- *Response variable* **Y is the one of which variation we want to explain** (*dependent variable*)
- Background variables such as elevation and temperature are *predictors* **X**, or *explanatory variables*, *independent variables,* **that we are using to explaing Y**
- *Simple linear regression* = one predictor
- *Multiple linear regression* = two or more predictors



Fitting a regression model

- In a simple regression model

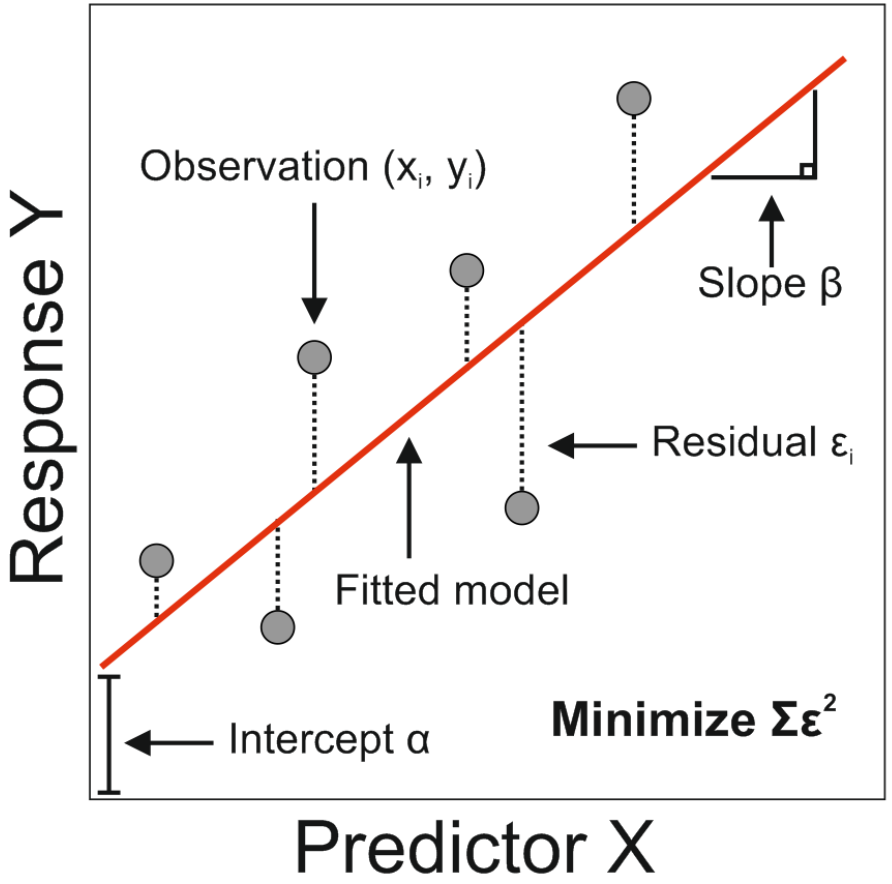
$$y = \alpha + \beta x,$$

we estimate two unknown *parameters* α, β where y_i is observed response variable and x_i is observed predictor i = 1,..,n. This can be written as:

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

where ε_i is *model residual*

- Model's parameters are being estimated using *ordinary least-squares* method, estimates are chosen to minimize the *sum of squared errors*
- Residual i.e. error term indicates the difference between the fitted model and observation

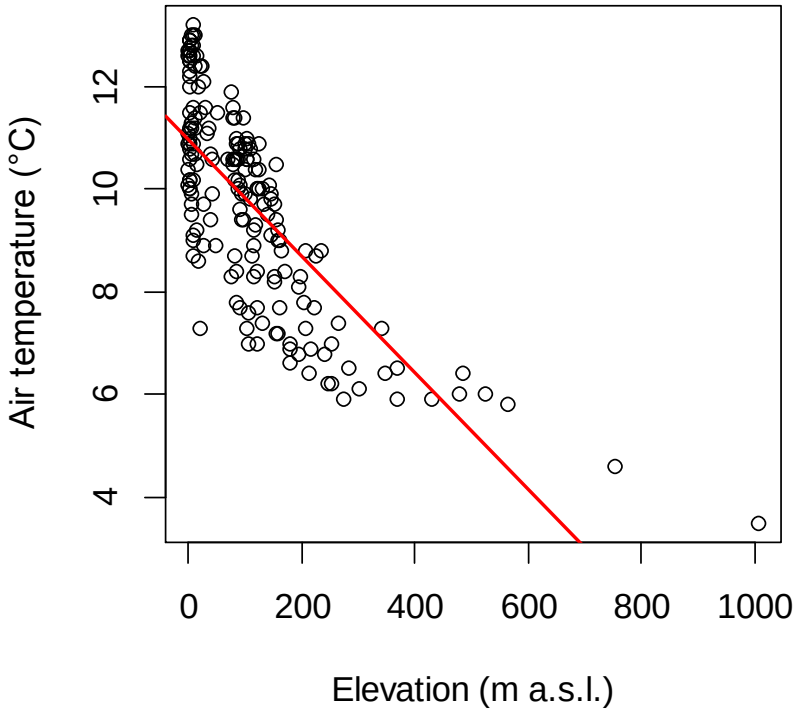


Interpretation of the model

- Consider a model where air temperatures (*y*) are being explained by elevation (*x*)
- **In regression model *y = a + bx*, the coefficient *b* indicates how much *y* changes on average, when *x* changes one unit**
- After fitting a linear regression model, the estimated *intercept* is 10.97 and *regression slope* is -0.01
- This relationship can now be written as:

Air temperature = 10.97 + (-0.01) x elevation

The model suggests that air temperature decrease, on average, by 0.01°C with increase in elevation by one unit (m)



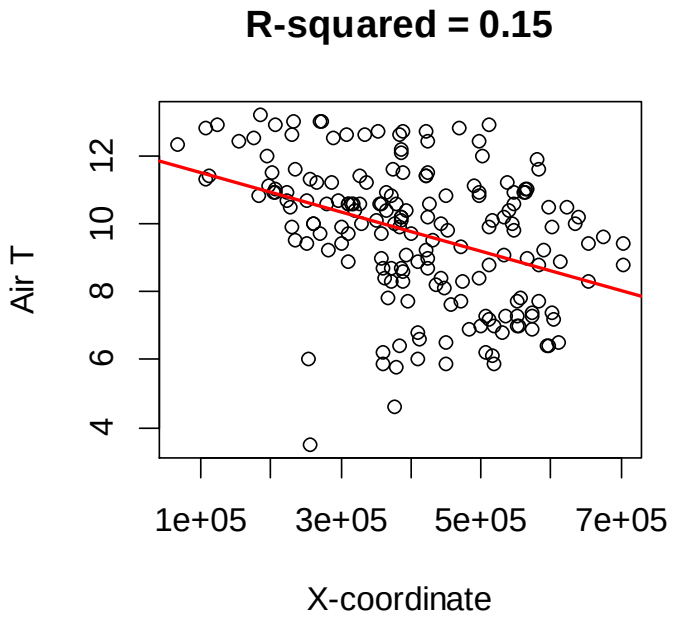
Interpretation of the model

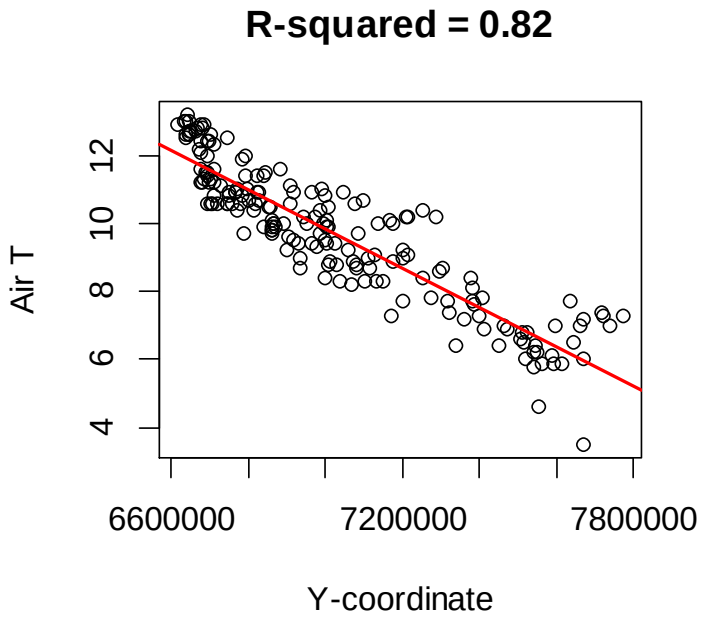
- Each estimated parameter is tested for *statistical significance* (*p-value*)
- Probability that the parameters' value is equal to zero (i.e. no effect)
- Low p-value (<0.05, *significant effect*), changes in elevation **are likely to be** related to air temperatures
- High p-value (>0.05, *insignificant effect*), changes in elevation **are not likely to be** related to air temperatures
- Standard error (Std. Error) tells you the precision of the estimate (*uncertainty*)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.9714859  0.1247025   87.98   <2e-16 ***
d$elev      -0.0113683  0.0007151  -15.90   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model diagnostics

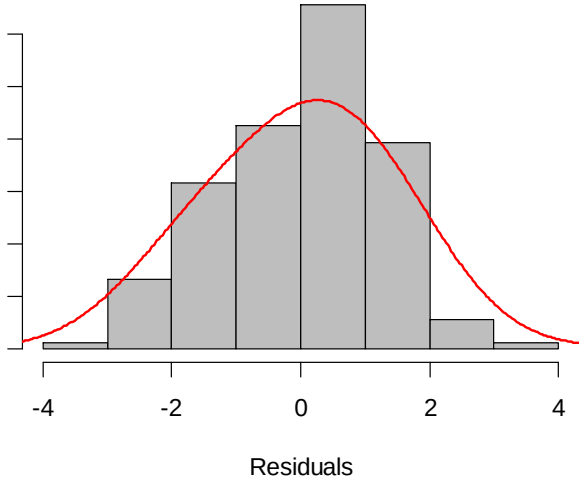
- How well the mathematical function (i.e. regression line) fits to the data?
- R-squared (R^2) indicates the amount of variation (0-1) of the response that is explained by the model
- Often expressed as percentages: 100 x R^2
- In simple linear regression, R^2 is the square of correlation coefficient R
- Suppose R^2 = 0.15, this means that 85% of the variability is still unaccounted for
- Note that R^2 increases (or stay constant) whenever you add new predictor to the model

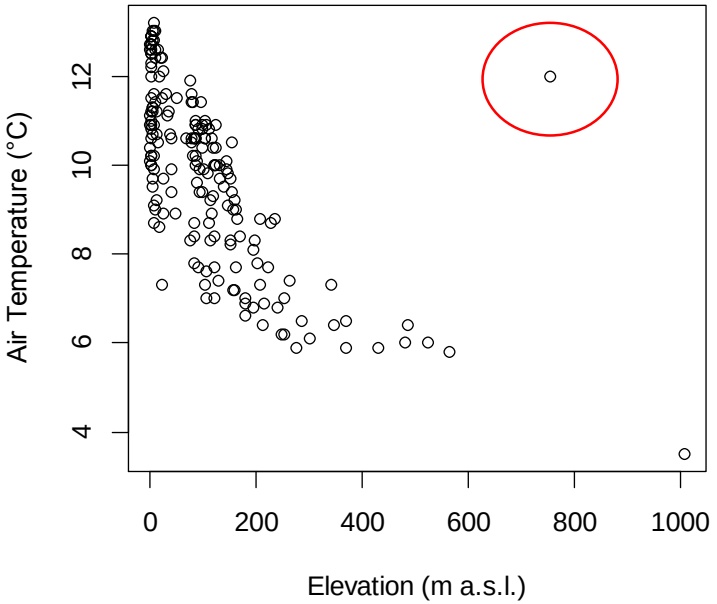




Key assumptions

- **Validity; data you are analyzing should be relevant to the research question you are trying to answer**
- Response and predictor are in a linear relationship (by their parameters!)
- Residuals ε are *normally distributed* (Gaussian)
- Predictors are uncorrelated (multicollinearity in multiple regression)
- Independence of residuals ε, no spatial autocorrelation
- Homoscedasticy = residuals have constant variance
- No major *outliers*



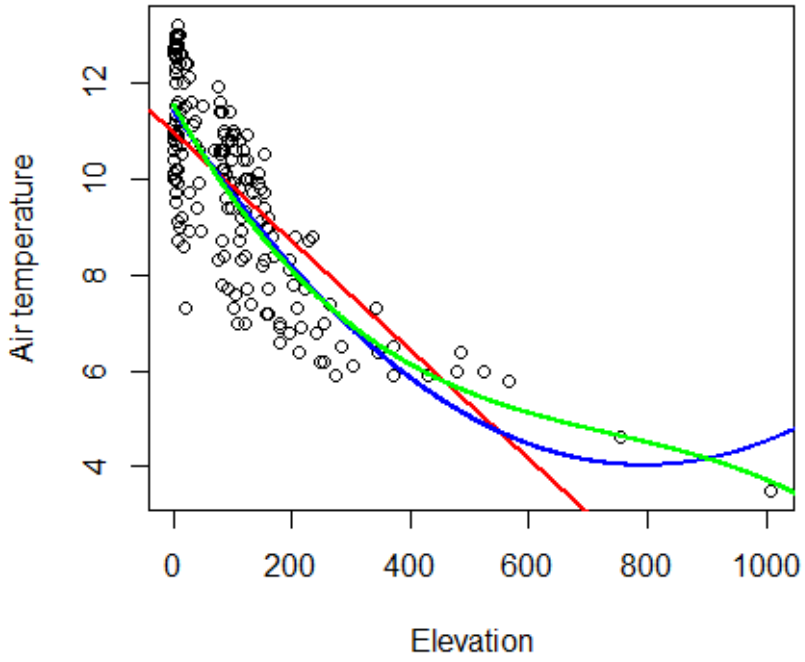


Higher-order polynomial terms

- Often response-predictor relationships do not follow a straight line
- Test for a *Curvilinear relationship* by including higher order polynomial term to the model

$$y = \alpha + \beta x + \mathbf{\beta x^2 + \beta x^3} + \varepsilon$$

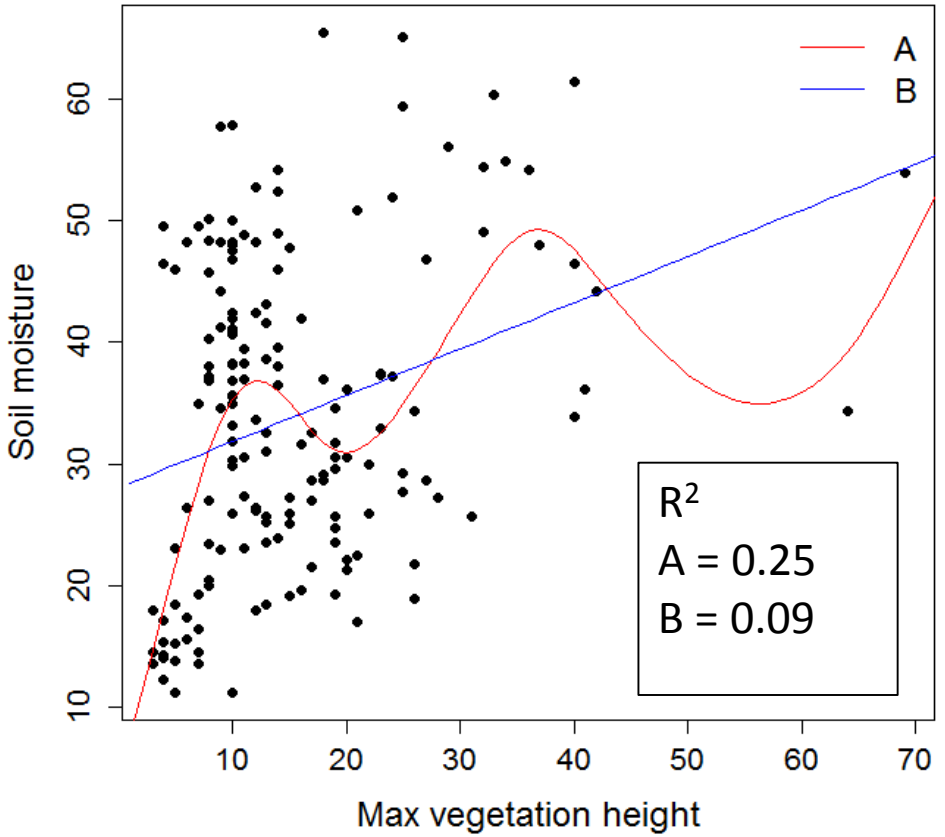
- This is still a linear model by its parameters!




```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.142e+01  1.356e-01  84.274  < 2e-16 ***
d$elev      -1.868e-02  1.358e-03 -13.757  < 2e-16 ***
I(d$elev^2)  1.179e-05  1.921e-06   6.138  5.2e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model transferability

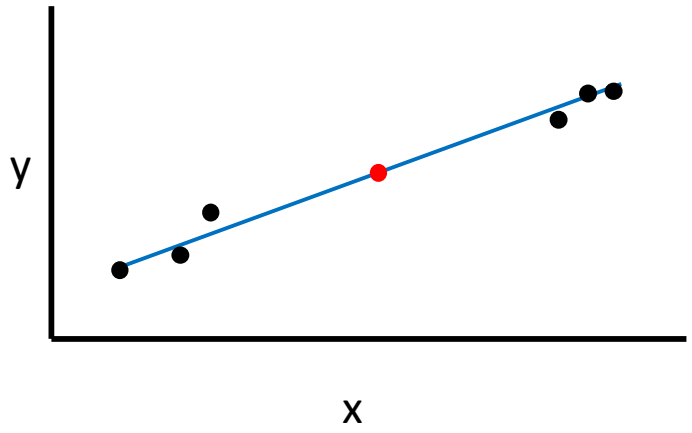
- Can model be generalized outside of it's data domain (e.g. another region)?
- Model fit can be high, but the real-life applicability low
- *Underfitting* vs. *overfitting* model



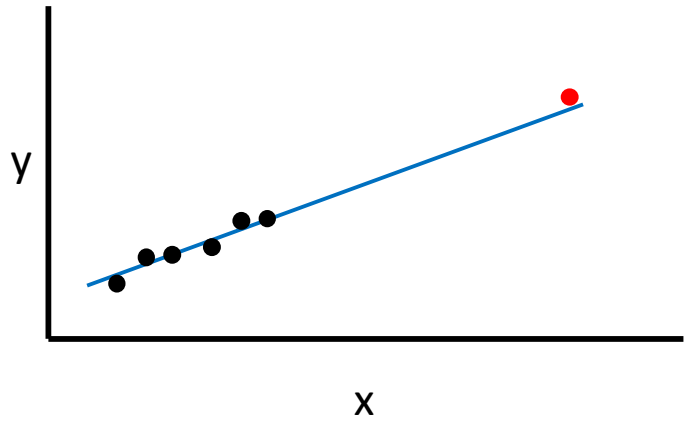
Linear regression – prediction

- A value of Y_i can be predicted, given that value of X_i is known

Interpolation: prediction **inside** data domain



Extrapolation: prediction **outside** data domain



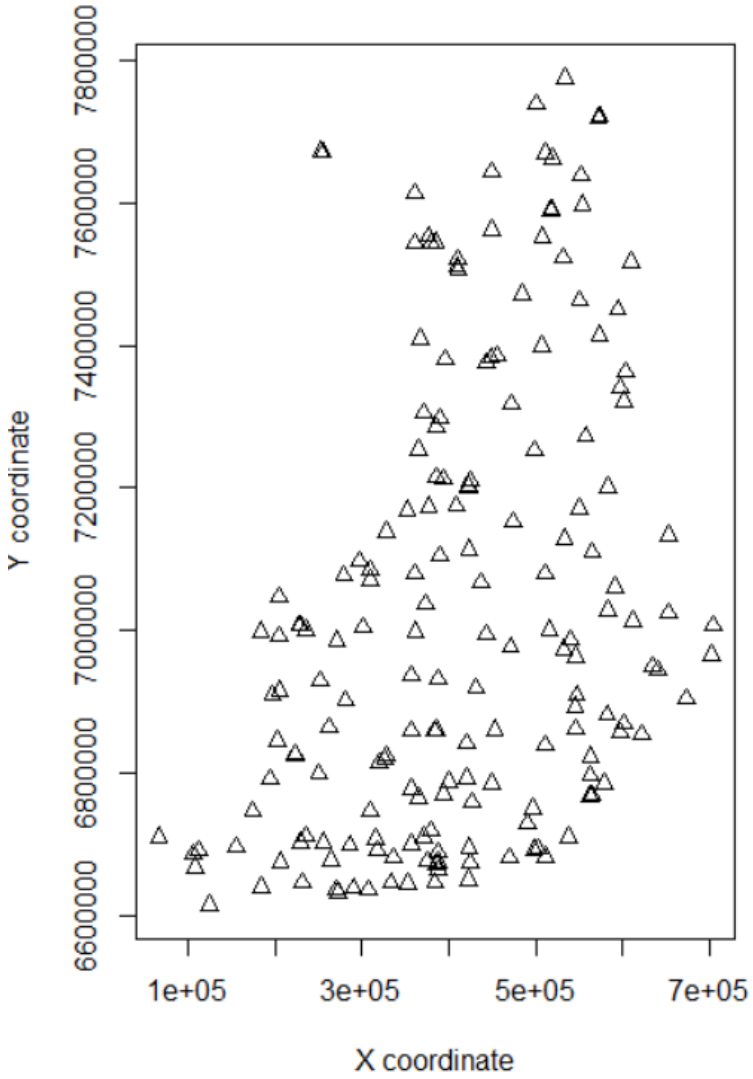
- Which of the models presented in the previous slide (A or B) is likely to produce unrealistic results when extrapolating outside the sample?

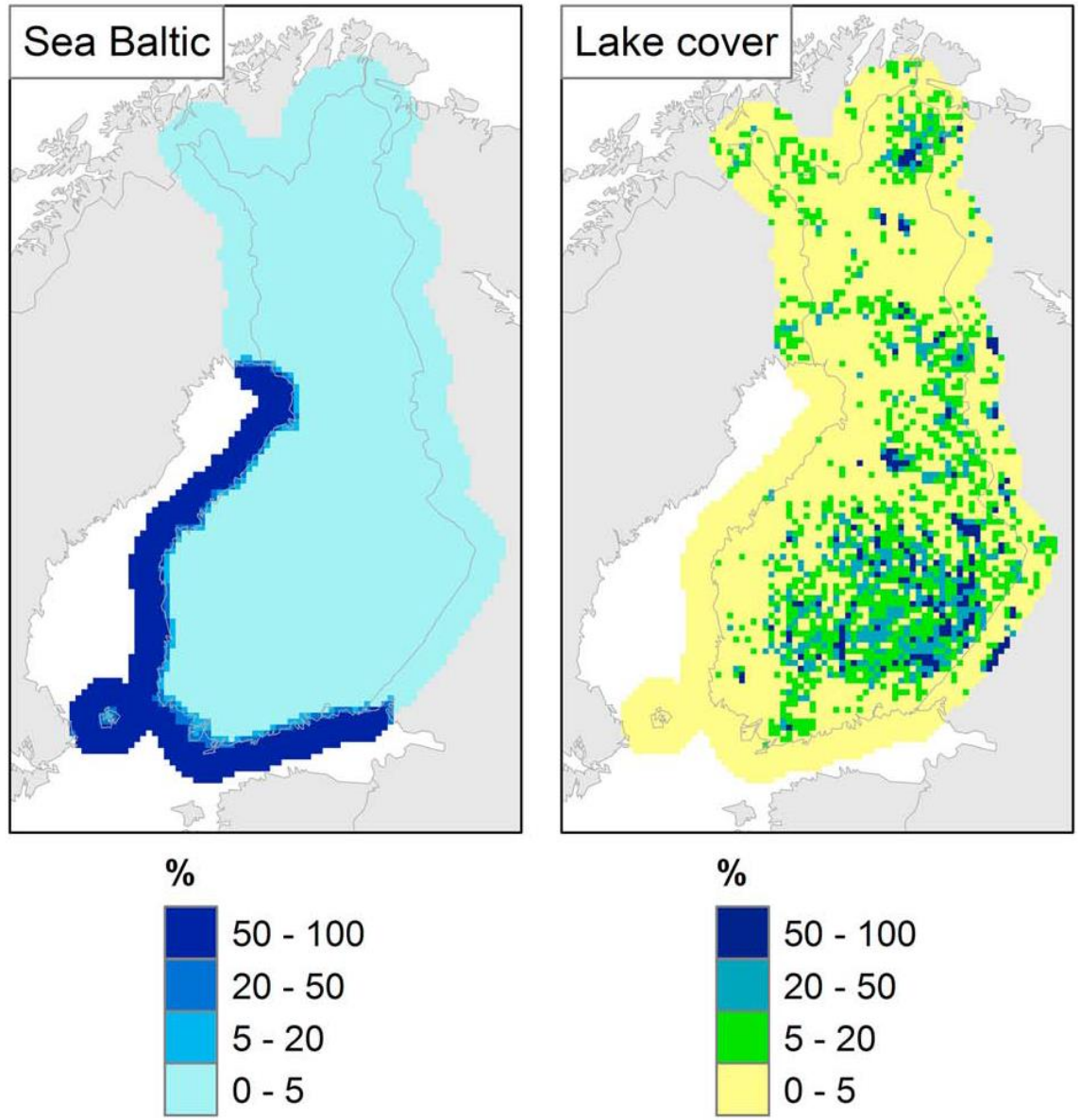
Linear regression in R – main functions

- *?lm*
- *lm()* for fitting a linear regression model
- *summary()* for extracting model results (estimated coefficients and their p-values)
- *anova()* significange of the model terms, model inter-comparison
- *abline()* draw the fitted line for simple regression to a plot
- *I(x^2)* define second order polynomial term
- *resid()* extract model residuals
- *fitted()* extract fitted values from the model object
- *predict()* use the fitted model to predict *Yi* at *Xi*

Practical – modeling monthly climate conditions in Finland

- R-script ”*LM_Practical1.pdf*”
- ”*AirTemperatureData.csv*”, average air temperatures
- *”PrecipitationData.csv”,* precipitation sums
- pvm = year and month
- station = FMI station id
- **temp/prec = air temperature (°C)/precipitation (mm)**
- x and y = geographical location (Euref fin)
- elev = elevation above sea level (m)
- lake = lake percentage (%), see next slide!
- sea = sea percentage (%), see next slide!




Sea Baltic
%
50 - 100
20 - 50
5 - 20
0 - 5
Lake cover
%
50 - 100
20 - 50
5 - 20
0 - 5

Questions

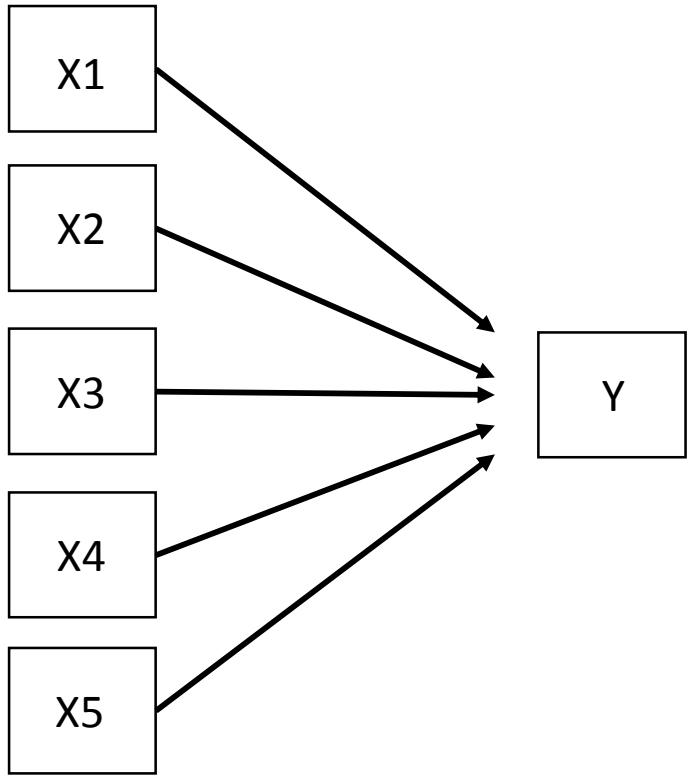
1. Based on the data ("*AirTemperatureData.csv*"), are air temperatures related to proximity to Baltic sea ("sea")? Is this relationship straight line or curvilinear? Plot both linear and quadratic response curves. ***p* < 0.05, F-test!**
2. Describe is the effect of latitude on average air temperatures? How much variation in air temperatures latitude explains?
3. According to our data, which one explains air temperatures better, lake predictor or longitude (x coordinate)? Plot response curves
4. Using July precipitation data ("*PrecipitationData.csv*") what single predictor explains the most variation in monthly precipitation sums, based on 1st and 2nd order polynomial terms?
5. Fit models of precipitation sum using (i) first order elevation term, and (ii) first and second order elevation terms. Which of the models is "better"? Based on the models, what are the predicted precipitation sums at 4500 m a.s.l.?

Multivariate models

Modelling in physical geography, 5cr, 30.10.-30.11.2017

Rationale

- Multivariate models refer to statistical techniques that can analyze the relationships of two or more variables at the same time
- Most of research questions in physical geography are multivariate by their nature
- In other words, to solve problems we need to consider multiple different predictors in the models
- Multivariate modelling can indicate the variables that explain the most variation in response
- It also can be used to control for the effects of other confounding predictors on a response variable



Rationale

- *Multivariate methods* is a general term that encompasses various functions, such as model identification, parameter testing, hypothesis testing, residual inspections, etc, …
- The common feature is that they simultaneously analyze multiple variables to model complex relationships in the data
- For example, water quality in river systems is controlled by many environmental factors:

  - soil conditions, bedrock, topography and land use

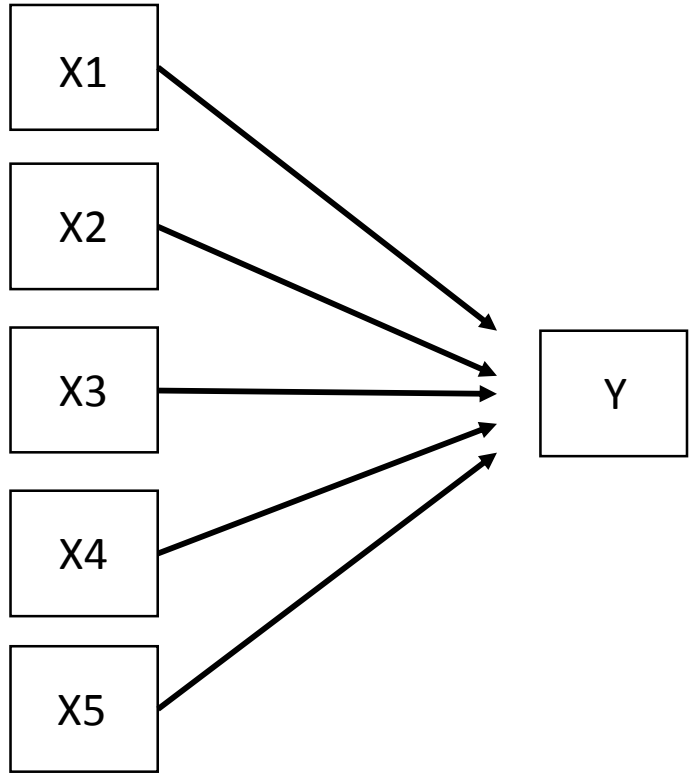
- Therefore multivariate methods are needed to understand the nature, effects and statistical properties of spatial phenomena
- Various statistical approaches: regression, classification trees, structural equation models, generalized boosting methods, etc…

Multiple regression model

- Simple regression can be extended to multiple regression:

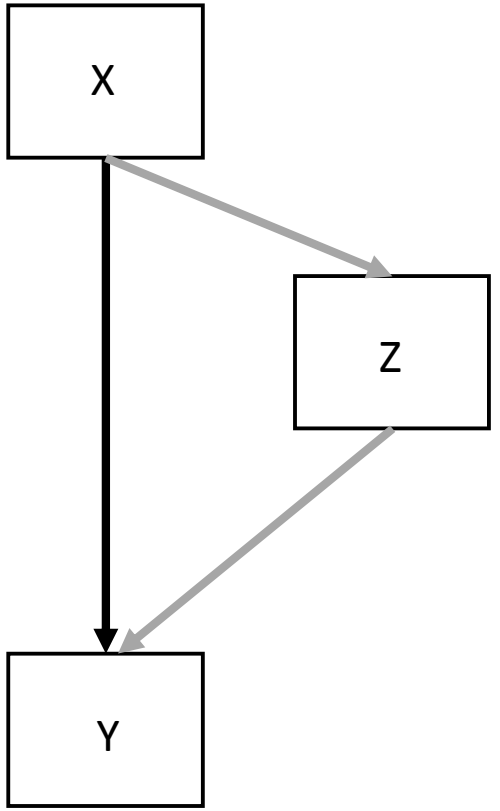
$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} \ldots \beta_k x_{ki} + \varepsilon_i,\ i = 1, \ldots, n$$

- The model can also consist of higher order polynomial terms and interactions
- Interpretation of the regression coefficients: **average change in *y*, when x_k changes one unit and <u>other terms are held constant</u>**



Interaction

- Statistical interaction means that we cannot assess variables main effects on response variable separately; instead we need to take account the effect of other predictor as well
- In another words, the **effect of *x* on *y* depends on the value of *z***
- Another variable can either strengthen or weaken (i.e. moderate) the main effect
- Example 1: moisture and nutrients have effect on plant grow through interaction; the effect of added water differs across nutrient conditions
- Example 2: the effect of incoming solar radiation on air temperatures varies across elevations



Interaction - interpretation

- Consider a model in R syntax, where species richness is explained using soil temperature (°C), soil moisture (% VWC) and their interaction

lm(richness ~ soil_temp*soil_moist, data=d)

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.169764   0.929065   9.870  < 2e-16 ***
soil_temp        -0.214344   0.115025  -1.863  0.06249 .
soil_moist        0.134726   0.044323   3.040  0.00239 **
soil_temp:soil_moist  0.012063   0.005923   2.037  0.04175 *
---
```

1st coefficient = intercept; 2nd coefficient = effect of soil temperature on richness, when soil moisture = 0 % (not very meaningful …)

3rd coefficient = effect of soil moisture on richness, when soil temperature = 0°C

4th coefficient = If soil moisture increase by one unit, the regression slope between soil temperature and richness increase by 0.012

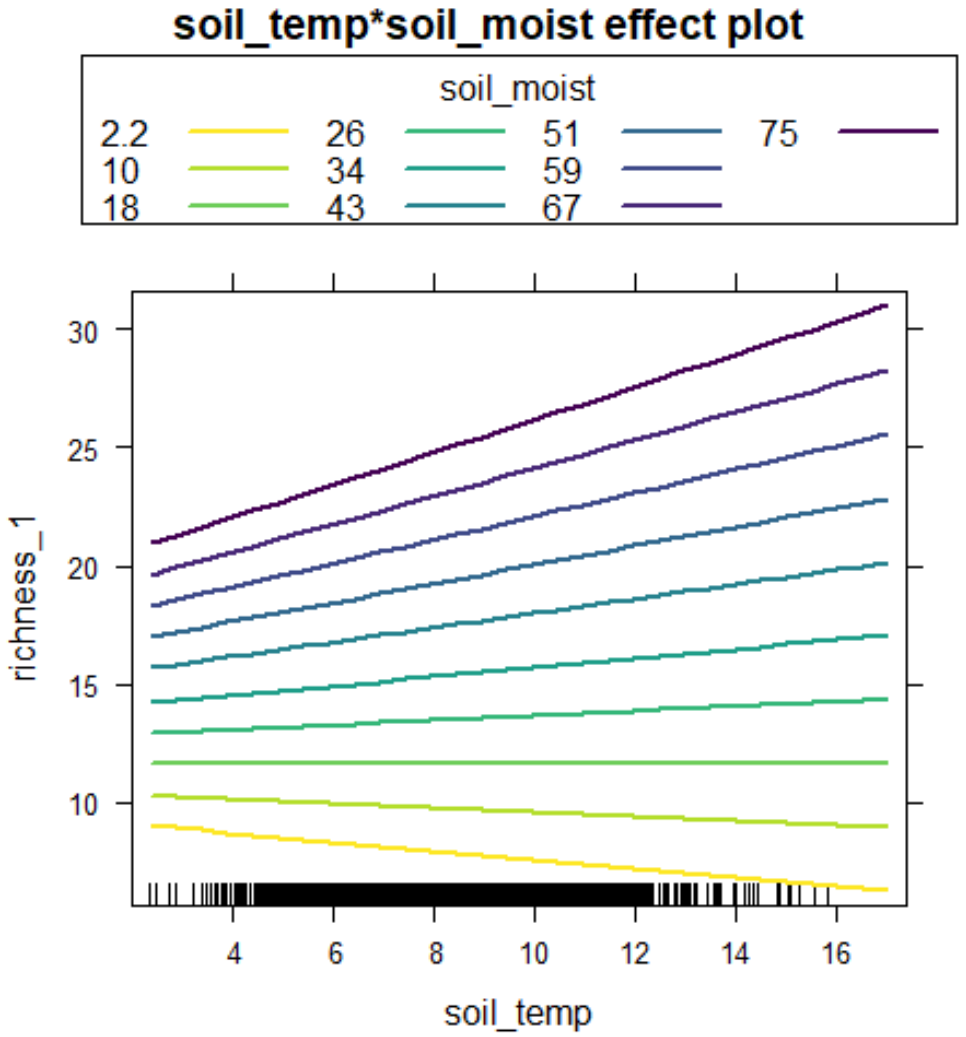
Interaction - interpretation

- Soil temperature has a strong positive effect on species richness at high soil moisture levels
- The effect changes from positive to negative at low soil moisture levels
- If no interaction, the effect of soil temperature would be the same over soil moisture levels

> require(effects); require(viridis)

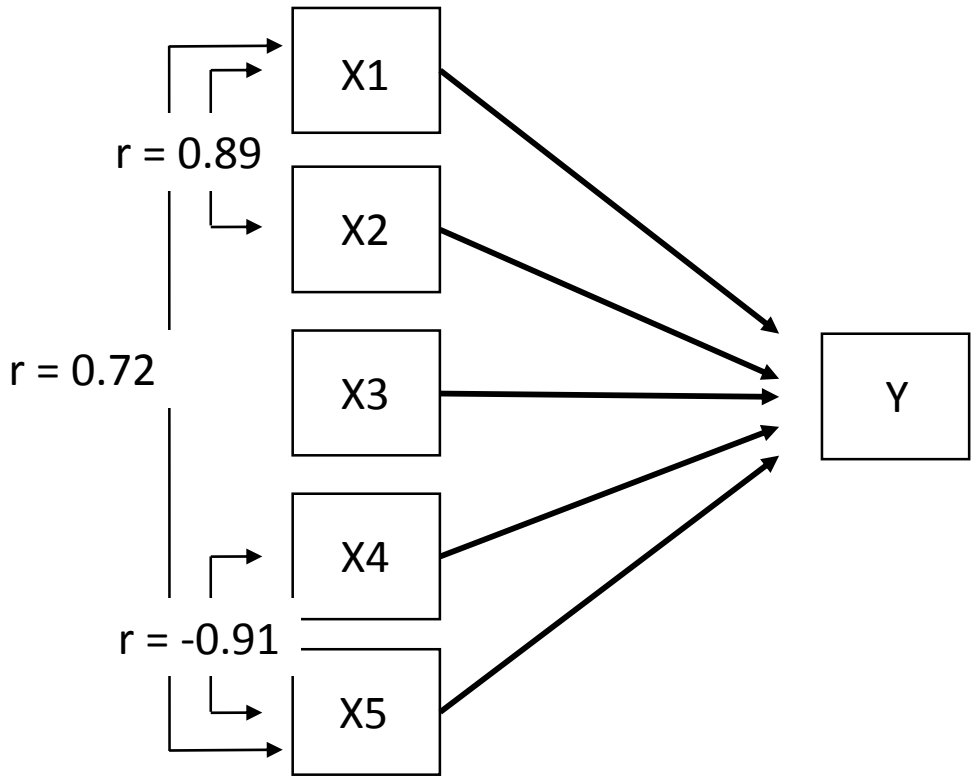
> m <-lm(richness_1~soil_temp*soil_moist, data=d)

> plot(effect(term="soil_temp:soil_moist", mod=m, xlevels=10), multiline=T, colors = rev(viridis(10)), lwd=2)



Variable selection

- Following a *principle of parsimony*, choose as few predictors as possible from a pool of candidate predictors, that explain the most of variation in response variable
- *Multicollinearity* among predictors potentially confounds the found effects
- Rule of thumb: consider only predictors with pairwise r<[0.7] in the models
- **Choose predictors based on known (direct) relationships, well established statistical criteria and/or prevailing theory! You have to be able to justify your model**



Stepwise variable selection

- Backward stepwise variable selection; start with a *full model* consisting of all potential predictors, and omit each term at a time having the highest *p-value* (i.e. predictor that is unlikely to have effect on response)
- Continue until you are left only with significant terms (e.g. $p \leq 0.05$)
- Consider a regression model

  $y = \alpha + \beta_1 x + \beta_2 z + \beta_3 x^2 + \beta_4 z^2 + \beta_5 x{:}z$

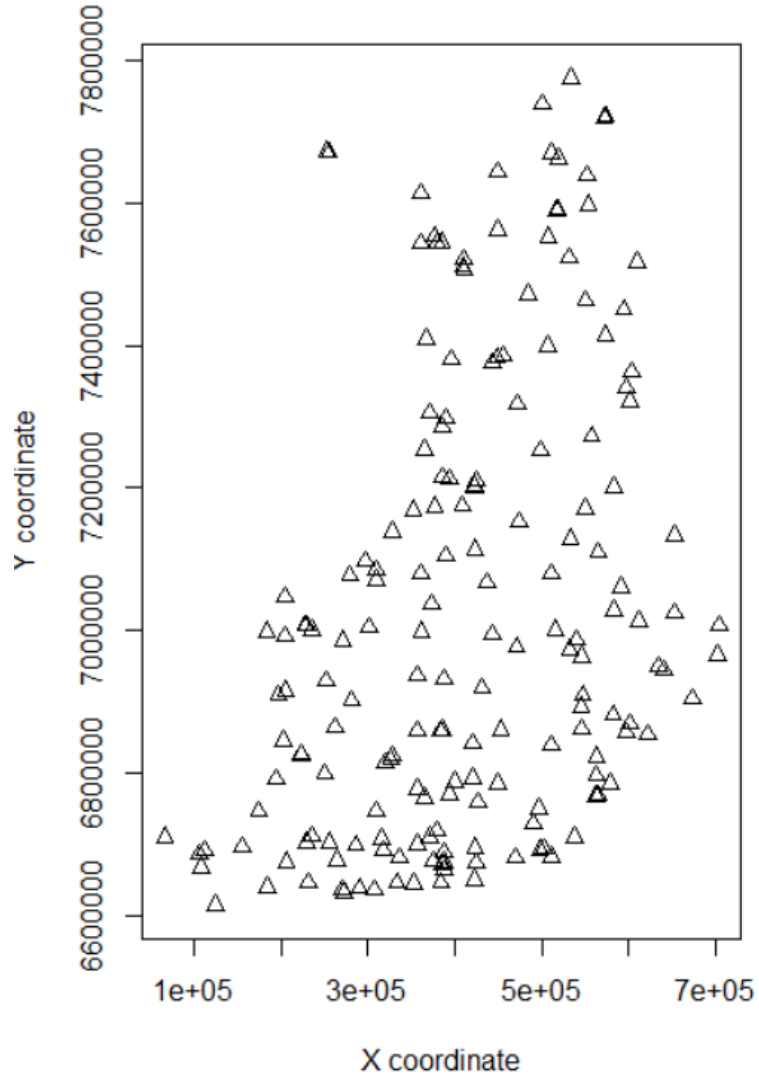
- At first step you can consider omitting terms $\beta_3 x^2$, $\beta_4 z^2$ or $\beta_5 x{:}z$
- We can consider omitting first order terms only after higher order terms and interaction has been omitted
- If model has term x^2 it must also include x

Stepwise variable selection

- Three approaches for stepwise variable selection; *forward*, *backward* and their *combination* (both ways)
- Manual variable selection may not be convenient, if modelling studying multiple response variables or when modelling multiple time steps
- Automated variable selection procedures are implemented in R. For example MASS::stepAIC() for variable selection based on Akaike's Information Criteria (AIC)
- Note: *p*-values and other statistical criteria for stepwise variable selection are indicative, and other properties such as *effect size*, together with theory should be preferred

Practical – factors influencing monthly average air temperatures in Finland

- R-script ”*LM_Practical2.pdf*”
- ”*AirTemperatureData.csv*”, average air temperatures
- pvm = year and month
- station = FMI station id
- **temp = air temperature (°C)**
- x and y = geographical location (Euref fin)
- elev = elevation above sea level (m)
- lake = lake percentage (%), see next slide!
- sea = sea percentage (%), see next slide!



Questions

1. Based on our data ("*AirTemperatureData.csv*"), what are single most influential variables (elevation, lake, sea, x and y) explaining the spatial variation in average September air temperatures in Finland? Are the effects (i) linear or curvilinear and (ii) positive or negative? ***p* < 0.05, F-test!**
2. What is the effect of elevation on air temperatures, after the effect of other environmental factors is controlled (based on first order polynomial terms)? How would you interpret the results?
3. What is the effect of lake predictor of air temperatures after the effect of other environmental factors is controlled (based on first order terms)? How much of the variation in air temperatures this model explains?
4. Find the most parsimonious model explaining average air temperatures using backward stepwise variable selection (based on first and second order terms, define interaction term for geographical location [x and y]). Consider for multicollinearity among predictors (r<[0.7]).