Swiss TPH

Swiss Tropical and Public Health Institute
Schweizerisches Tropen- und Public Health-Institut
Institut Tropical et de Santé Publique Suisse

Assoziiertes Institut der Universität Basel

Workshop on epidemiological analysis of air pollution effects on vegetation Basel, September 16/17, 2014

Statistical methods for epidemiological analysis

Dr. C. Schindler
Swiss Tropical and Public Health Institute
University of Basel
christian.schindler@unibas.ch

Contents

Causality and Confounding

Random effects models

Measure of model fit and model selection criteria

Modeling of non-linear functional relationships

Causality and confounding

Causality criteria by Austin Bradford Hill (1965)

http://www.edwardtufte.com/tufte/hill http://en.wikipedia.org/wiki/Austin_Bradford_Hill

These criteria are generally neither necessary nor sufficient but provide guidance.

Clarification of terminology

Exposure: Any factor / variable that might causally influence our outcome variable of interest (e.g., ozon concentration, temperature, soil composition, etc.)

Disease indicator: a) Manifest tree disease or
b) Symptom indicating some problems
c) Quantitative parameter indicating some problems (e.g., decreased growth)

Association: Association between disease indicator of interest and exposure(s) of interest observed in a given study.
Notice that associations can be positive (more disease at higher exposures) or negative (less disease at higher exposures)

Bradford-Hill’s 9 criteria

1. Strength: strength of association adds credibility to a causal relationship

2. Consistency: association repeatedly observed in different studies / settings rules out chance as explanation and makes it unlikely that a common systematic error is at work.

3. Specificity: The exposure in question is asscoiated with a specific disease type but not with other disease types. No other factors (strongly) increasing the risk of the respective disease are known.

4. Temporality: The exposure must have been present (for enough time) before the disease occurred / the disease indicator increased.

5. Biological gradient: Presence of a dose-response relation: Increase / decrease of risk of disease / disease indicator with increasing level of exposure.

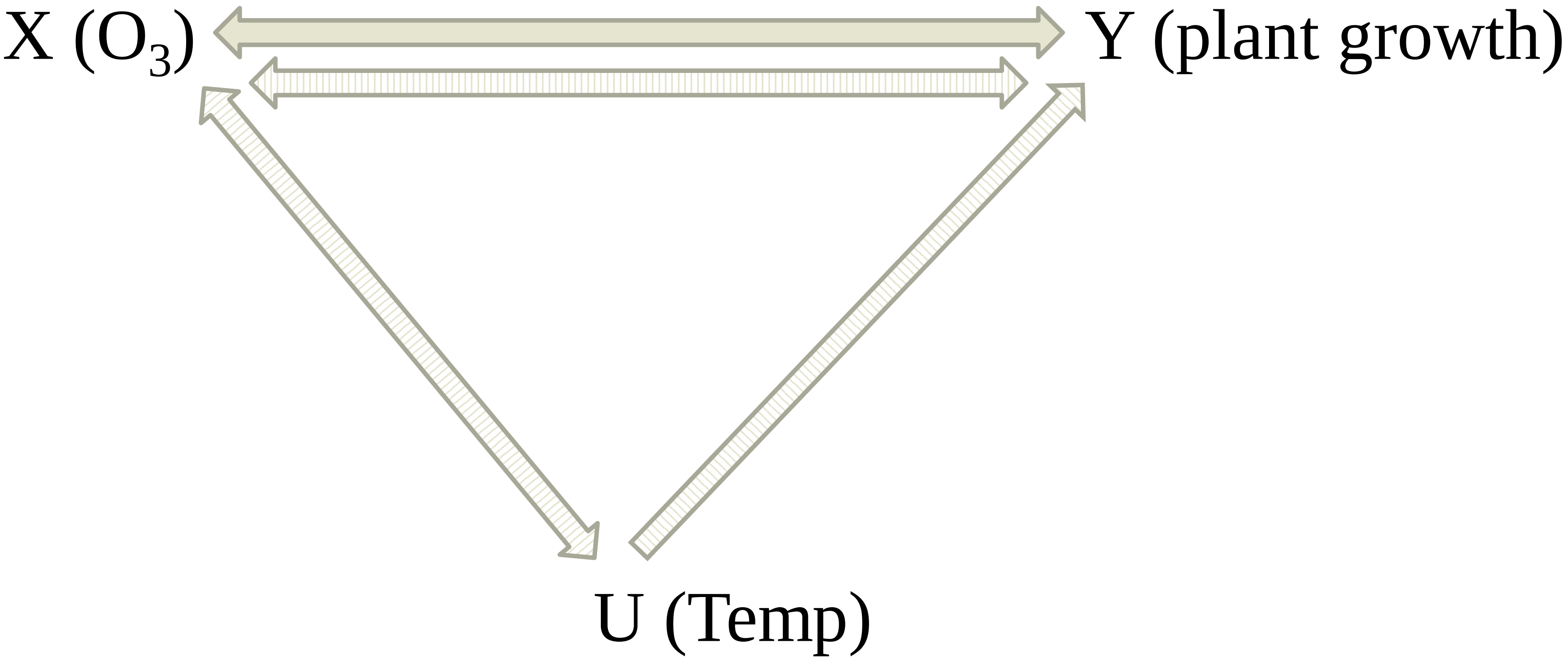
6. Biological plausibility: The association makes sense in the light of current biological knowledge.

7. coherence: Observed association should not conflict with existing evidence on the respective disease, in particular with established associations between the disease and other factors.

8. experiment: If association can be demonstrated under experimental conditions, this is the strongest evidence for its causality

9. Analogy: Were similar associations found for similar disease indicators and/or exposures in the past?

Confounding



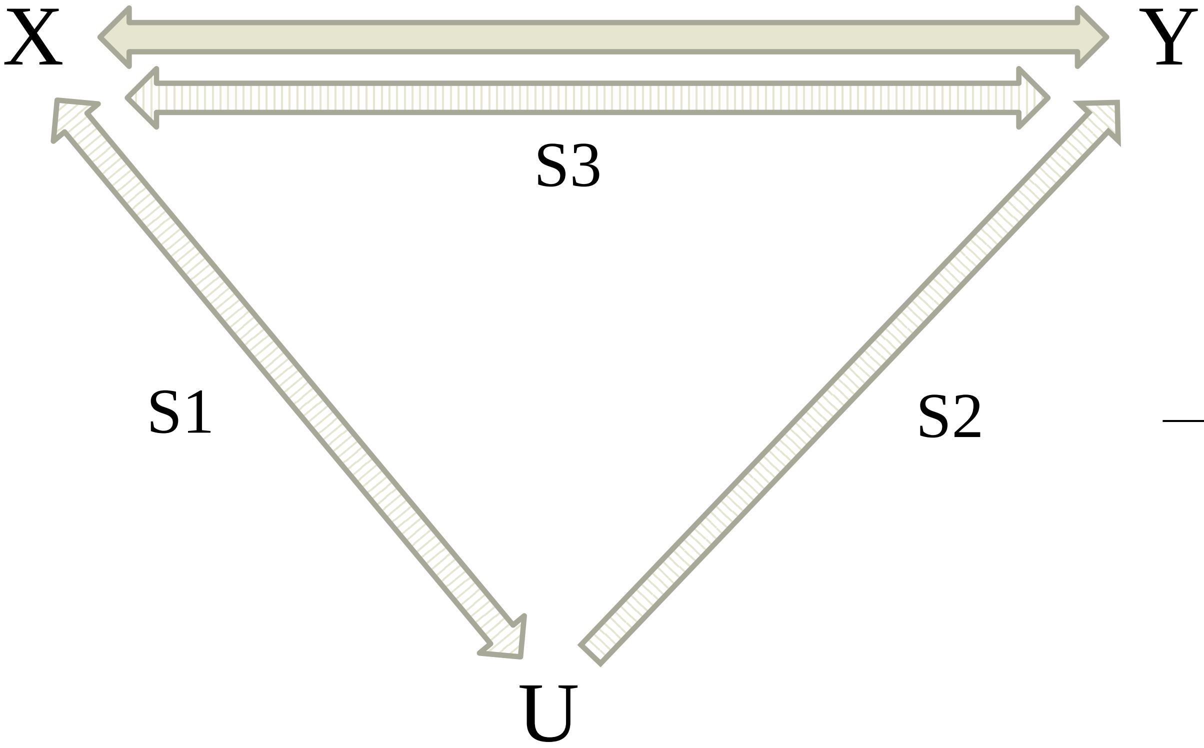
X (O_3)

Y (plant growth)

U (Temp)

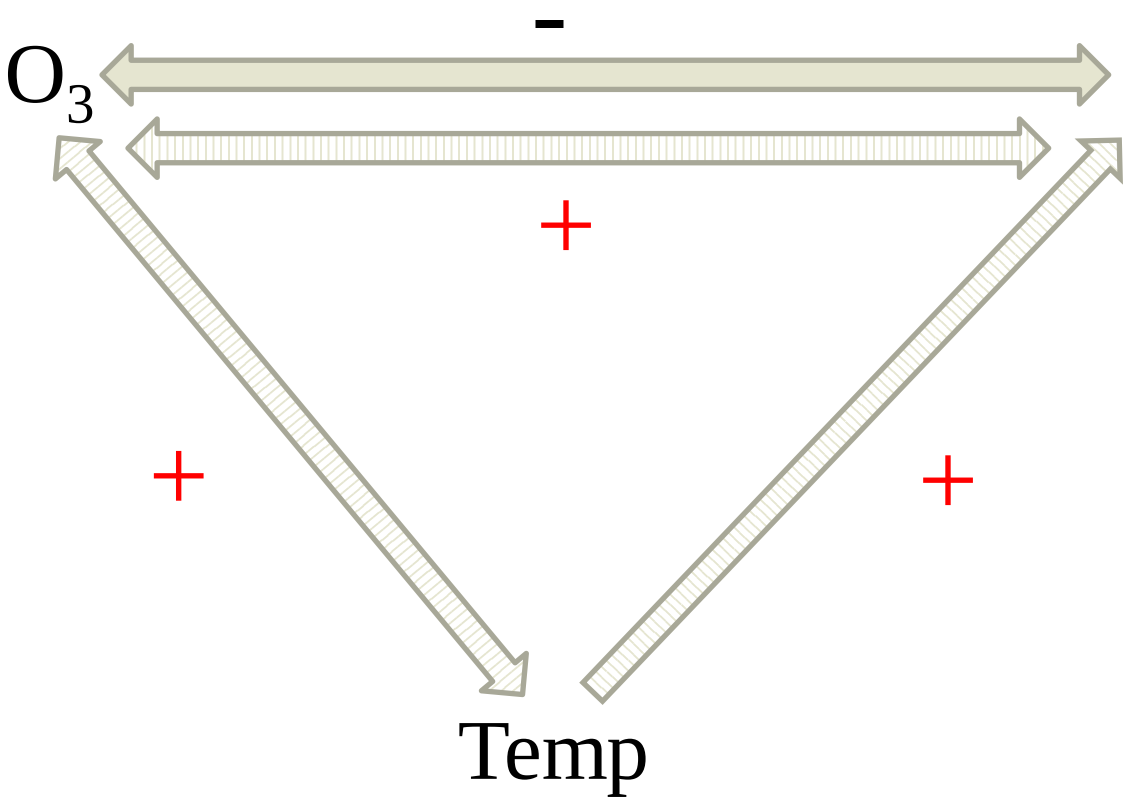
Bidirectional arrows indicate association, one-directional arrows causal influence.

Arithmetic of confounding



S1	S2	S3
+	+	+
+	-	-
-	+	-
-	-	+

Example: O_3, Temp and plant growth

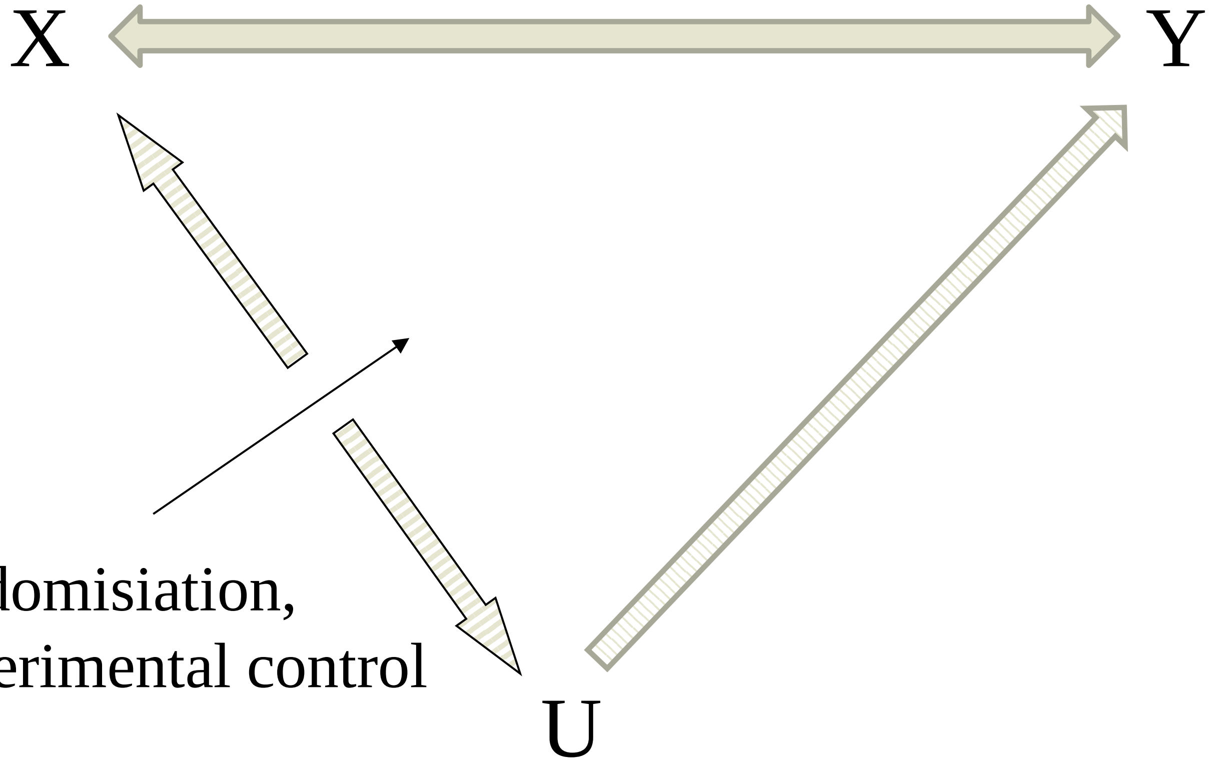


S1	S2	S3
+	+	+
+	-	-
-	+	-
-	-	+

Negativity of association between plant growth and O_3 is underestimated if effects of temperature are not controlled for in the model.

How can confounding be avoided or at least minimized?

1. Experimental blocking of confounder



randomisiation,
experimental control

2. Stratifikation by confounder variable

Example: Assess relation between plant growth and ozone separately in areas with higher and lower temperature levels.

3. Regression models

$$Y \;=\; b_0 \;+\; b_1 \cdot X \;+\; b_2 \cdot U \;+\; \text{«random» influence*}$$

Influence of X on Y

Influence of U on Y

or more generally:

$$Y \;=\; f(X, U, \text{random influence})$$

* includes influences from unmeasured factors

Residual confounding

Typically occurs in stratified analyses, because the confounder may still show some variation within strata.

May also occur in regression models if the effect(s) of the confounder are not well modelled or if the confounder is not measured precisely enough.

What must be considered when modeling the effect(s) of a confounder?

1. Potential non-linearities in the effect.
   If effects are modeled as linear despite their non-linearity, this may lead to RC*.

2. Interactions of the confounder with other factors.
   Their ignoring may also lead to RC*.

3. If the metric chosen to measure the confounder is not appropriate, this may lead to RC*.

4. If the lag structure of effects is not properly reflected in the model, this may lead to RC*.

* whenever the variable of interest can step in to mitigate the respective modeling deficit.

Consequences of model mis-specification

For a prediction model, the consequences may be minor if the model is always applied under the conditions having been present when the model was derived.

But these underlying conditions are likely to be different in other geographic regions and they also tend to change over time in the same region.

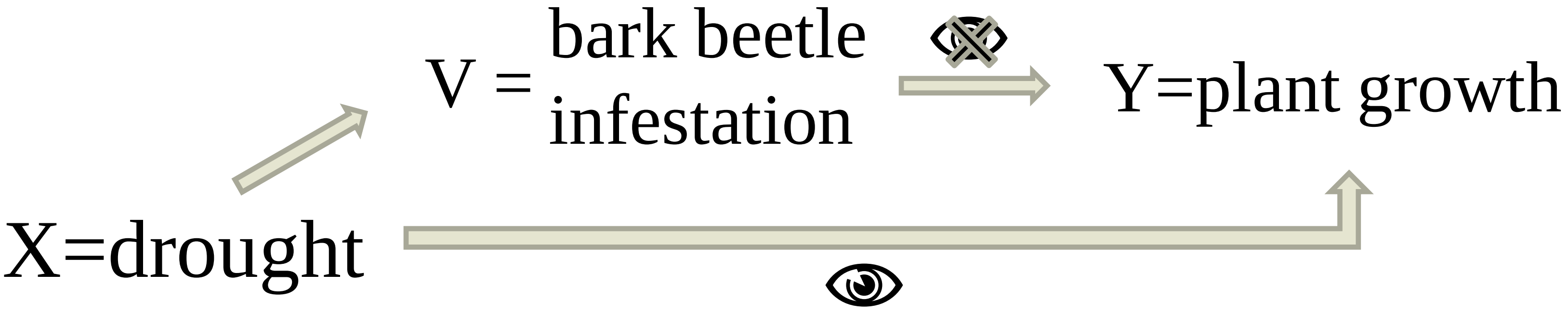
Therefore prediction models should be extrapolated to other regions with caution and they should be regulary updated even for their «native» use.

Effect estimates tend to be biased in the presence of residual confounding,
i.e., they tend to be systematically wrong.

Cave: **intermediate endpoints**

Intermediate endpoints V are parameters or events on the pathway from the exposure of interest X to the endpoint of interest Y. They are often wrongly treated as confounders of X. Their inclusion in the model absorbs effects from X on Y which are mediated by V. As a consequence, only the effects of X on Y which are not mediated by V are observed.

Example:



V = bark beetle infestation

Y=plant growth

X=drought

How to deal with intermediate endpoints?

may

a) omit them

or

b) regress V on X and replace V by the residuals of this regression.

or

c) use structural equation models to disentangle the effects of X on Y mediated by V and those not going through V.

Random effects models

Random effects models

$$Y = b_0 + b_1 \cdot X_1 + \ldots + b_p \cdot X_p + R_{andom}$$

In classical linear models we assume that different observational units share no common random influences, i.e., that

$$R = R_{measurement}$$

In random effects models, R_{andom} is considered to be the sum of random influences from more than one source, e.g.,

$$R_{andom} = R_{region} + R_{plot} + R_{tree} + R_{year} + R_{measurement}$$

Generally, repeated measurements from the same tree or
measurments from trees in the same plot
or from plots in the same region
share common random influences

These random influences are nested:
individual measurements are nested in trees
trees are nested in plots
plots are nested in regions

On the other hand, random effects of the factor year capture specific influences in the different years unexplained by the predictor variables X_1, X_2, …, X_p of the model. They are frequently assumed to be identical across all regions.

Region, plot, tree and year are referred to as **cluster levels** and the members of a cluster level are referred to as **clusters**.

Members of the cluster level «region» are the different regions
‘’ of the cluster level «plot» are the different plots in
the respective region
‘’ of the cluster level «tree» are the different trees in
the respective plot

Mixed linear models and generalized linear mixed models can deal with such random effects.

They will not estimate them as they estimate the parameters b_1, b_2, …., b_p of the predictor variables X_1, X_2,…,X_p, instead they estimate the variances of the random effects.

And they assume that each random effect is an outcome of a normal distribution with mean 0.

$R_{measurement}$ <- $N(0, \sigma_e^2)$

R_{tree} <- $N(0, \sigma_{tree}^2)$

R_{plot} <- $N(0, \sigma_{plot}^2)$

R_{region} <- $N(0, \sigma_{region}^2)$

R_{year} <- $N(0, \sigma_{year}^2)$

What happens if random effects are ignored?

Then the parameter estimates, i.e., the estimates of the coefficients b_1, b_2, …., b_p of the predictor variables

X_1, X_2, …., X_p may be biased

(but this needs not be the case)

However, the standard errors, p-values and confidence intervals will inevitably be biased.
Statistical significance of factors varying (mainly) between clusters is overestimated, while statistical significance of factors varying (mainly) within clusters is underestimated.

Why not replace random effects by fixed effects?

This is always an option, especially if the number of clusters is limited (e.g., in case of a limited number of regions).

But this generally leads to a loss of statistical power because contrasts in a predictor variable X (e.g., O_3) between the clusters are no longer available for estimating the effect of X.

Moreover, there may be cluster-level variables (variables that do not vary within but only between clusters, e.g., altitude of the plot) that one would like to include among the predictors. This is impossible if the respective clusters are represented in the model by a fixed factor.

Measures of model fit

and

model selection criteria

Measures of model fit in classical regression

R^2 Proportion of the variance of Y (outcome) which the model explains in the underlying sample

adjusted R^2 Estimate of the proportion of the variance of Y which the model explains at the population level

$$= R^2 - (1 - R^2) \cdot p / (n - 1 - p) \quad *$$

The adjusted R^2 is to be preferred because R^2 increases with each additional variable even if this variable is uninformative for Y.

For the adjusted R^2, larger is better.

* n = sample size, p = number of parameters other than cons

Measures of model fit in generalized linear models

AIC Akaike information criterion

$$\underbrace{-\ 2*\ln(\text{likelihood})}_{\text{measure of misfit in sample}} \quad + \quad \underbrace{2 \cdot \text{number of parameters*}}_{\substack{\text{penalty for model complexity} \\ \text{(risk of overfit)}}}$$

BIC Bayes information criterion

$$\underbrace{-\ 2*\ln(\text{likelihood})}_{\text{measure of misfit in sample}} \quad + \quad \underbrace{\ln(n) \cdot \text{number of parameters*}}_{\text{penalty for model complexity}}$$

For both measures smaller is better

* other than cons

AIC and BIC in classical linear models

AIC Akaike information criterion

$$\underbrace{\text{n*ln(variance of residuals)}}_{\text{measure of misfit in sample}} + \underbrace{2 \cdot \text{number of parameters*}}_{\substack{\text{penalty for model complexity} \\ \text{(risk of overfit)}}}$$

BIC Bayes information criterion

$$\underbrace{\text{n*ln(variance of residuals)}}_{\text{measure of misfit in sample}} + \underbrace{\text{ln(n)} \cdot \text{number of parameters*}}_{\text{penalty for model complexity}}$$

For both measures smaller is better

* other than cons

Model cross-validation and AIC

To get an idea of how well a model will be able to predict new observations, **cross-validation** may be used.

Leave-one-out cross-validation works as follows:

a) The first observation of the derivation sample of the model is omitted.

b) The model is refitted without this observation, providing model M(-1)

c) The difference between y of the omitted observation and its prediction from M(-1) is computed and stored.

d) The procedure a) – c) is repeated with the rest of the observations.

e) The variance of all obtained differences is computed (cross-validation error variance)

Among different competing models, the one with the smallest error variance is selected. However, AIC may be used instead, because it is very closely related to the cross-validation error variance.

When to use which criterion?

Adjusted R^2 and BIC are good criteria for comparing classical linear prediction models, if the aim is to only include truly informative variables. Each predictor variable should have a clearly visible influence on adj R^2.

AIC is less strict than BIC and may be preferred in explanatory models where the aim is to include both truly informative and potentially informative variables.

In any case, model comparison based on BIC or AIC is superior to model comparison based on p-values, since no significance level can be justifed by theoretical considerations. For instance $\alpha = 0.05$ is pure convention without any theoretical underpinning.

Modeling of non-linear functional relationships

Modeling of non-linear functional relations

1. Use of polynomial functions

Recommended: prior centering of predictor variables (e.g., by subtracting the mean)

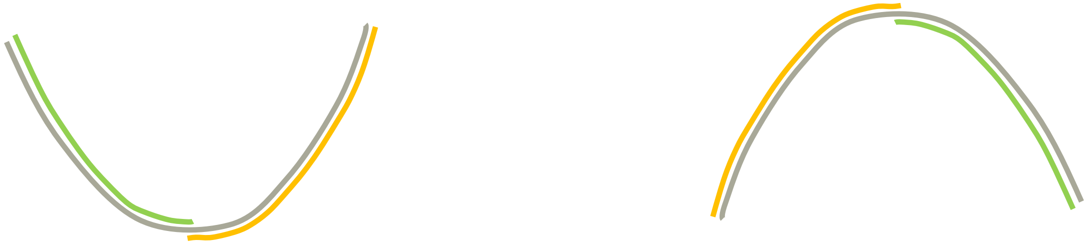
x1c <- x1 – mean(x1)

model statement: y ~ x1c + I(x1c^2) + I(x1c^3) + x2c + ….

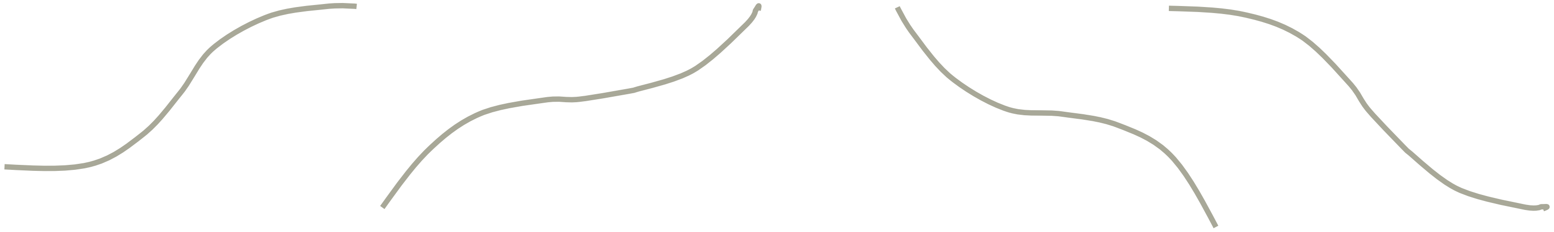
or

model statement: y ~ poly(x1c,3) + x2c + ….

Quadratic functions can model functional relationships of the form

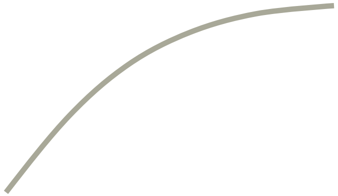


Cubic functions can model functional relationships of the form



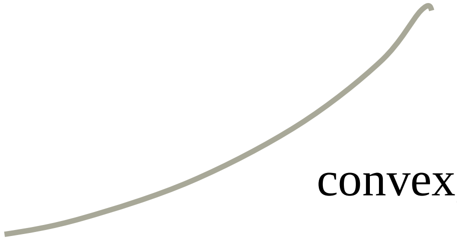
2. Use of fractional polynomials

Use of functions $x^3 \quad x^2 \quad x \quad \sqrt{x} \quad \ln(x) \quad {}^{1}\!/_{\sqrt{x}} \quad {}^{1}\!/_{x} \quad {}^{1}\!/_{x^2}$



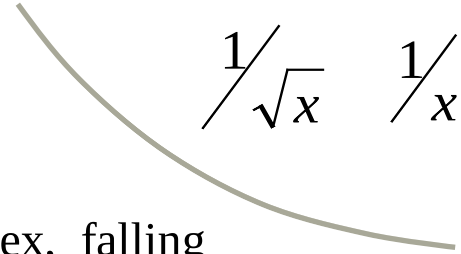
concave, rising

$\sqrt{x} \quad \ln(x)$



$x^2 \quad x^3$

convex, rising



${}^{1}\!/_{\sqrt{x}} \quad {}^{1}\!/_{x} \quad {}^{1}\!/_{x^2}$

convex, falling

R-syntax

R-package: mfp

```
fpmod <- mfp( y ~  fp(x1) + fp(x2) + x3 + …. , data=file)
```

Effects of variables wrapped by fp() are modelled as fractional polynomials

print(fpmod) generates model output

To plot the functional relationship between y and x1

```
filec <- file

filec$x2 < - mean(x2)

filec$x3 < - #
```

(Generate new data frame where all predictor variables other than the one to be plotted are set to their mean value or a certain fixed value)

```
pred<-predict(model,newdata=filec,level=0)
```

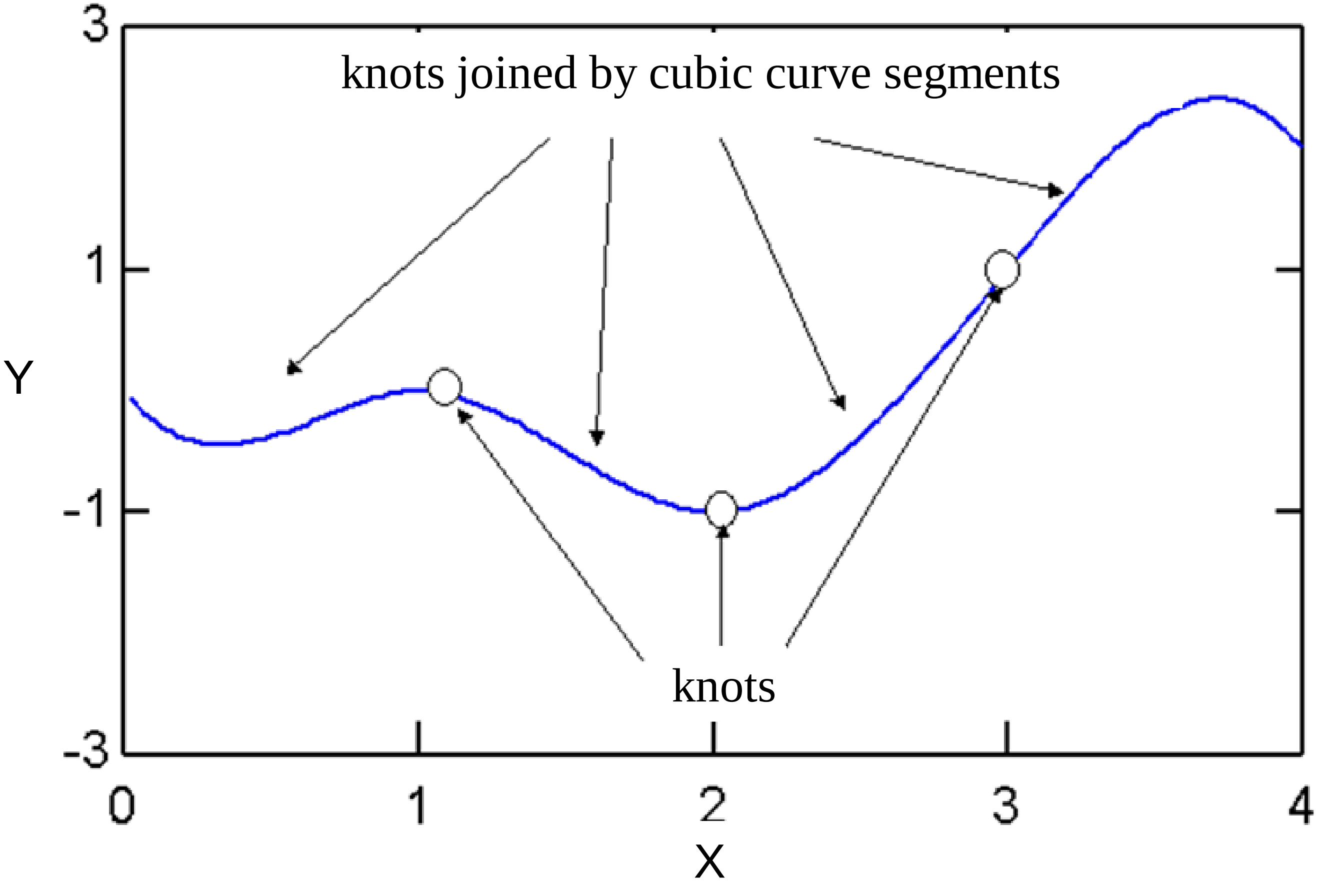
(Evaluate the prediction equation at the values of file.)

```
plot(file$x1,pred)
```

(Plot the curve between pred and x1)

3. Use of splines

Some splines are built by joining different curve segments in a smooth way:



Cubic spline

Natural splines

Natural cubic splines are cubic splines, as in the previous picture, but with linear end pieces.

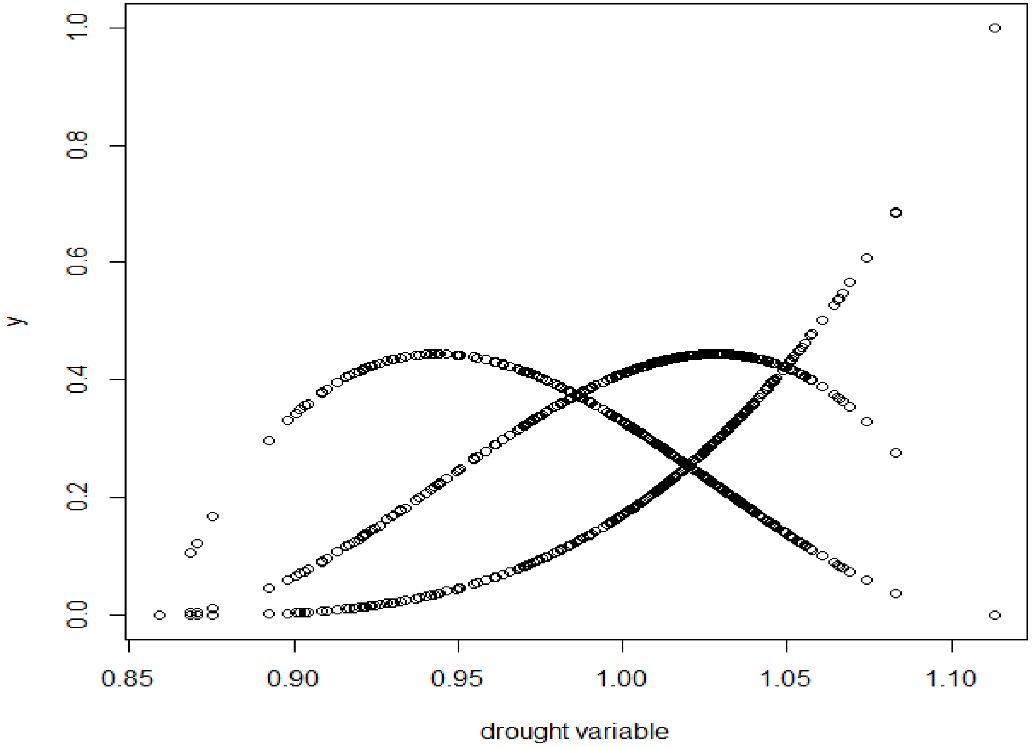
Compared to cubic splines they are less prone to produce artefacts at the ends of the domain of X.

Linear splines are automatically natural, because they are composed of linear pieces over consecutive intervals.

Quadratic splines are composed of quadratic curve segments and it they are natural, they also have linear end pieces.

B-splines

are superpositions of basis functions each of which is zero outside a fixed interval



Cubic B-spline basis functions defined over the range of the drought variable

B-splines are very flexible and collinearity problems are not an issue with them.

Natural splines and B-splines in R

R-Package: splines   Natural splines (ns) and B-splines

```
model <- lme(y ~ ns(x1,df=#) + x2 + x3, random = ~1|site,  data=file)
```

The «optimal» number of degrees of freedom can be determined using the AIC or the BIC-criterion

```
AIC(model) or BIC(model)
```

To plot the spline function:

Can proceed as with the fractional polynomials, but predict command must be extended to tell R what to do with the random effects.

```
pred<-predict(model,newdata=filec,level=0)
```

(Evaluate the prediction equation at the values of file.)

```
plot(file$x1,pred)
```